Deep learning for prediction of colorectal cancer outcome: a discovery and validation study

Elsevier Editorial System(tm) for The Lancet
Manuscript Draft

Manuscript Number: THELANCET-D-19-03766R2

Title: Deep learning for prediction of colorectal cancer outcome: a discovery and validation study

Article Type: Article

Corresponding Author: Professor Havard E Danielsen, Ph. D.

Corresponding Author's Institution: Oslo University Hospital

First Author: Ole-Johan Skrede, M. Sc.

Order of Authors: Ole-Johan Skrede, M. Sc.; Sepp De Raedt, Ph. D.; Andreas Kleppe, Ph. D.; Tarjei S Hveem, PhD; Knut Liestøl, Ph. D.; John Maddison, Ph. D.; Hanne A Askautrud, Ph. D.; Manohar Pradhan, M. D., Ph. D.; John Arne Nesheim, M. Sc.; Fritz Albregtsen, Ph. D.; Inger Nina Farstad, M. D., Ph. D.; Enric Domingo, Ph. D.; David N Church, D. Phil.; Arild Nesbakken, M. D., Ph. D.; Neil A Shepherd, D.M.; Ian Tomlinson, Ph. D.; Rachel Kerr, M. D., Ph. D.; Marco Novelli, M. D., Ph. D.; David J Kerr, M. D., D. Sc.; Havard E Danielsen, Ph. D.

Abstract: Background: Improved markers of prognosis are needed to stratify patients with early-stage colorectal cancer to refine selection of adjuvant therapy. The aim of the present study was to develop a biomarker of patient outcome after primary colorectal cancer resection by directly analysing scanned conventional haematoxylin and eosin stained sections using deep learning.
Methods: More than 12,000,000 image tiles from 828 patients with distinctly good or poor disease outcome were used to train a total of 10 convolutional neural networks, purpose-built for classifying supersized heterogeneous images. A prognostic biomarker integrating the 10 networks were determined using 1645 patients with non-distinct outcome. The marker was tested on 920 patients with slides prepared in UK, and finally independently validated according to a pre-defined protocol in 1122 patients treated with single-agent capecitabine using slides prepared in Norway. The primary outcome was cancer-specific survival.
Findings: The biomarker provided a hazard ratio for poor vs good prognosis of 3·84 (95% confidence interval, 2·72-5·43; p<0·0001) in the primary analysis of the validation cohort, and 3·04 (95% confidence interval, 2·07-4·47; p<0·0001) after adjusting for established prognostic markers significant in univariable analyses of the same cohort; pN stage, pT stage, lymphatic invasion, and venous vascular invasion.
Interpretation: It was possible to develop a clinically useful prognostic marker using deep learning allied to digital scanning of conventional haematoxylin and eosin stained tumour tissue sections. The assay has been extensively evaluated in large, independent patient populations, correlates with and outperforms established molecular and morphological prognostic markers, and gives consistent results across tumour and nodal stage. The biomarker stratified stage II and III patients into sufficiently distinct prognostic groups that these potentially could be used to guide selection of adjuvant treatment by avoiding therapy in very low risk groups and identifying patients who would benefit from more intensive regimes.

Deep learning for prediction of colorectal cancer outcome: a discovery and validation study

Ole-Johan Skrede, M. Sc.[1,2,*], Sepp De Raedt, Ph. D.[1,2,*], Andreas Kleppe, Ph. D.[1,2], Tarjei S. Hveem, Ph. D.[1], Prof. Knut Liestøl, Ph. D.[1,2], John Maddison, Ph. D.[1], Hanne A. Askautrud, Ph. D.[1], Manohar Pradhan, Ph. D.[1], John Arne Nesheim, M. Sc.[1], Prof. Fritz Albregtsen, M. Sc.[1,2], Prof. Inger Nina Farstad, Ph. D.[3,4], Enric Domingo, Ph. D.[5], David N. Church, D. Phil.[6,7], Prof. Arild Nesbakken, Ph. D.[4,8,9], Prof. Neil A. Shepherd, D. M.[10], Prof. Ian Tomlinson, Ph. D.[1,11], Prof. Rachel Kerr, Ph. D.[5], Prof. Marco Novelli, Ph. D.[1,12], Prof. David J. Kerr, D. Sc.[13], Prof. Håvard E. Danielsen, Ph. D.[1,2,13]**

[1]Institute for Cancer Genetics and Informatics, Oslo University Hospital, Oslo, Norway

[2]Department of Informatics, University of Oslo, Oslo, Norway

[3]Department of Pathology, Division of Laboratory Medicine, Oslo University Hospital, Oslo, Norway

[4]Institute of Clinical Medicine, University of Oslo, Oslo, Norway

[5]Department of Oncology, University of Oxford, Oxford, UK

[6]NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, UK

[7]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK

[8]Department of Gastrointestinal Surgery, Oslo University Hospital, Oslo, Norway

[9]K.G. Jebsen colorectal cancer research centre, Oslo, Norway

[10]Gloucestershire Cellular Pathology Laboratory, Cheltenham General Hospital, Cheltenham, UK

[11]Edinburgh Cancer Research Centre, University of Edinburgh, Edinburgh, Scotland

[12]Research Department of Pathology, University College London Medical School, London, UK

[13]Nuffield Division of Clinical Laboratory Sciences, University of Oxford, Oxford, UK

*Both authors contributed equally to this work.

**Corresponding author:

Prof Håvard E. Danielsen,

Institute for Cancer Genetics and Informatics,

Oslo University Hospital

Montebello, 0310, Oslo, Norway

Email: hdaniels@labmed.uio.no

Phone: +47 22782320

Words in abstract (not exceed 300): 297

Words in main text (up to 3500): 3889

Number of references (up to 30): 30

Number of figures: 2

Number of tables: 3

Background: Improved markers of prognosis are needed to stratify patients with early-stage colorectal cancer to refine selection of adjuvant therapy. The aim of the present study was to develop a biomarker of patient outcome after primary colorectal cancer resection by directly analysing scanned conventional haematoxylin and eosin stained sections using deep learning.

Methods: More than 12,000,000 image tiles from 828 patients with distinctly good or poor disease outcome were used to train a total of 10 convolutional neural networks, purpose-built for classifying supersized heterogeneous images. A prognostic biomarker integrating the 10 networks were determined using 1645 patients with non-distinct outcome. The marker was tested on 920 patients with slides prepared in UK, and finally independently validated according to a pre-defined protocol in 1122 patients treated with single-agent capecitabine using slides prepared in Norway. The primary outcome was cancer-specific survival.

Findings: The biomarker provided a hazard ratio for poor *vs* good prognosis of 3·84 (95% confidence interval, 2·72-5·43; p<0·0001) in the primary analysis of the validation cohort, and 3·04 (95% confidence interval, 2·07-4·47; p<0·0001) after adjusting for established prognostic markers significant in univariable analyses of the same cohort; pN stage, pT stage, lymphatic invasion, and venous vascular invasion.

Interpretation: It was possible to develop a clinically useful prognostic marker using deep learning allied to digital scanning of conventional haematoxylin and eosin stained tumour tissue sections. The assay has been extensively evaluated in large, independent patient populations, correlates with and outperforms established molecular and morphological prognostic markers, and gives consistent results across tumour and nodal stage. The biomarker stratified stage II and III patients into sufficiently distinct prognostic groups that these potentially could be used to guide selection of adjuvant treatment by avoiding therapy in very low risk groups and identifying patients who would benefit from more intensive regimes.

Funding: The Research Council of Norway through its IKTPLUSS Lighthouse program (grant number 259204, project name DoMore!).

Research in context

Evidence before this study

Digital image analysis is one of the fields where the recent renaissance of deep learning has achieved the most impressive results. We searched PubMed on June 12, 2019 without language or time restrictions, using the terms “deep learning”, “prediction”, “survival”, “cancer”, and “histology” (full specification of the search criteria is provided in the appendix p 3). We systematically reviewed the 214 search results, and found 18 original research studies which applied deep learning to predict patient outcome or related attributes using histopathology images.

In 16 studies, the patient outcome was indirectly predicted by identifying attributes known to correlate with patient outcome, e.g. stromal fraction, mitotic count, or Gleason pattern. Two studies reported on direct prediction of survival, but neither presented a marker for automatic prediction of patient outcome from scanned whole-slide sections; one required manual annotation to locate interesting tissue regions, and the other classified tissue microarray spots. Perhaps even more importantly, neither of these two studies evaluated their biomarker in independent cohorts; the performance was instead estimated using cross-validation in the same cohort as utilised for training, which can easily lead to overoptimistic estimates.

Added value of the study

We have applied deep learning to develop a biomarker for automatic prediction of cancer-specific survival directly from scanned haematoxylin and eosin stained, formalin-fixed, paraffin-embedded tumour tissue sections. Independent validation demonstrated that the

biomarker improved prediction of cancer-specific survival by stratifying stage II and III colorectal cancer patients into distinct prognostic groups, supplementing established prognostic markers, and outperforming most existing markers in terms of hazard ratios. The marker could potentially be used to improve selection of adjuvant treatment after resection of colorectal cancer by identifying patients at very low risk who may have been cured by surgery alone, as well as patients at high risk who are much more likely to benefit from more intensive regimes.

Implications of all the available evidence

It is possible to utilise deep learning to develop biomarkers for automatic prediction of patient outcome directly from conventional histopathology images. In colorectal cancer, the marker was found to be a clinically useful prognostic marker in analysis of a large series of patients who received consistent, modern cancer treatment.

Introduction

Biomarkers are being used increasingly to match anticancer therapy to specific tumour genotypes, protein, and RNA expression profiles, usually in patients with advanced disease.[1–3] One example of this is selection of *KRAS*-wild-type colorectal cancers (CRCs) for treatment with epidermal growth factor receptor inhibitors.[4] However, in the adjuvant setting for CRC, the primary question is binary, whether to offer treatment at all, and subsequent selection of drugs, dose, and schedule is predominantly driven by stage rather than by companion diagnostics. If it were possible to further refine prognostic models, this could allow a more targeted approach by defining subgroups in which the absolute benefits of adjuvant chemotherapy are minimal, relative to surgery alone, and at the other end of the spectrum, patients who might benefit from prolonged combination chemotherapy because of their poor survival rate.[5–8]

More than two decades of adjuvant trials in patients with early-stage CRC using fluoropyrimidines, in combination with cytotoxic agents like oxaliplatin, have yielded an improved overall survival of around 3-5% for patients with stage II or IIIA CRC. Many patients are cured by surgery alone, while around 25% will recur despite adjuvant chemotherapy. There is likely to be a chemotherapy-associated death rate of 0·5-1%, and 20% of patients will suffer significant side-effects. The risk-benefit ratio is therefore rather marginal, but could potentially be much better if it were possible to define subgroups at higher or lower risk of recurrence and cancer-specific death.[9–12]

Although clinically validated prognostic biomarkers would facilitate adjuvant therapeutic decisions, very few have been sufficiently robustly validated for routine clinical application. A case can be made for assessment of mismatch repair (MMR) status,[13,14] as patients with MMR-deficient tumours tend to have a good prognosis. We have recently reported that measurement of tumour cellular DNA content (ploidy) in combination with stromal fraction

can stratify stage II patients into very good, intermediate, and poor prognostic groups.[15] Interestingly, analysis of driver mutations and RNA signatures has shown them to be individually weak prognostic markers and unable to guide clinical decision making.[8,14] Deep learning refers to the class of machine learning methods that make use of successively more abstract representations of the input data to perform a specific task. These methods use a training set to learn how these representations should be generated in a manner appropriate for the given task. In contrast, traditional machine learning utilises handcrafted features to create representations of the input data that are applied to perform the task. In many applications, deep learning has been demonstrated to provide superior performance compared to other machine learning techniques, and it is a growing expectation that deep learning will transform current medical practice. Especially convolutional neural networks have excelled in many image interpretation tasks, and could therefore be hypothesised to retrieve additional information from histopathology images. The aim of the present study was to use deep learning to analyse conventional whole-slide images (WSIs) in order to develop an automatic prognostic biomarker for patients resected for primary CRC. The marker was trained using 828 patients with distinct prognosis from four cohorts, fine-tuned using 1645 other patients from the same four cohorts, and tested on slides prepared at a different laboratory from 920 patients. Finally, the marker was independently validated according to the pre-defined protocol (appendix pp 52-80) on 1122 patients analysed retrospectively from a trial (QUASAR 2) of adjuvant therapy.[16]

Methods

Training and Tuning Cohorts

Four different cohorts were utilised for training and tuning to achieve a broad patient representation and thereby improve the ability to generalise to new cohorts. Three cohorts

were consecutive series of stage I, II or III tumours from CRC patients treated at hospitals with both rural and urban catchment areas: (i) 160 patients treated 1988-2000 at Akershus University Hospital, Norway;[17] (ii) 576 patients treated 1993-2003 at Aker University Hospital, Norway;[15] and (iii) 970 patients treated in Gloucester 1988-1996 and included in the Gloucester Colorectal Cancer Study, UK.[18,19] The fourth cohort were 767 stage II or III CRC patients treated at 151 UK hospitals in 2002-2004 and included in the VICTOR trial (ISRCTN registry number ISRCTN98278138).[20] Our cohorts included only patients with resectable tumour, and a formalin-fixed, paraffin-embedded (FFPE) tumour tissue block available for analysis.

To obtain clear ground-truth, we used as training cohort the 828 patients with so-called distinct outcome, either good or poor. A patient was assigned to the good outcome group if aged less than 85 years at surgery, had more than six years follow-up after surgery, and had no record of recurrence or cancer-specific death. The poor outcome group consisted of those aged less than 85 years at surgery and suffered cancer-specific death between 100 days (inclusive) and 2·5 years (exclusive) after surgery. Patients not satisfying either of these group criteria were defined as having non-distinct outcome, and these 1645 patients were used for tuning. The protocol specifies additional cohort details, and demographics are summarised in table 1.

Test Cohort

The test cohort consisted of 920 patients from the Gloucester Colorectal Cancer Study, UK.[18,19] WSIs were obtained from different FFPE tumour tissue blocks than those used in the training and tuning cohorts.

Validation Cohort

The validation cohort consisted of 1122 patients from 170 hospitals in seven countries recruited to the QUASAR 2 trial (ISRCTN registry number ISRCTN45133151).[16] Inclusion

criteria were age 18 years or older, CRC adenocarcinoma histologically proven to be R0 M0 stage III or high-risk stage II, primary resection 4-10 weeks before randomisation, WHO performance status score 0 or 1, and life expectancy (with comorbidities, but excluding cancer risk) of at least five years. See protocol pp 22-25 for exclusion criteria and other details. All patients received adjuvant therapy, either capecitabine plus bevacizumab or capecitabine alone, with equal disease-free and overall survival in both trial arms.[16]

Sample Preparation

Slides in VICTOR cohort were prepared in Oxford, UK, while the other slides in the training and tuning cohorts were prepared at the Institute for Cancer Genetics and Informatics (ICGI), Norway. Introducing this variation in the development phase was hypothesised to increase the robustness and generalisability of the trained marker. Slides in the test cohort were prepared as a part of the routine histopathological examination in Cheltenham, UK, and the performance in this cohort should thus indicate the prognostic ability when the marker is assayed at a different laboratory using original slides. Slides in the validation cohort were prepared at ICGI. All slides were made by staining a three µm FFPE tissue block section with haematoxylin and eosin (H&E), and a pathologist (MP) ascertained that it contained tumour. WSIs were acquired at the highest resolution available (referred to as 40x magnification by the manufacturers) on two scanners, an Aperio AT2 (Leica Biosystems, Germany) and a NanoZoomer XR (Hamamatsu Photonics, Japan).

Areas with high tumour content were identified using a segmentation network that was trained on a subset of the training and tuning cohorts (protocol pp 6-10). A WSI with the so-called 40x resolution typically contained an order of 100,000x100,000 pixels, multiple orders of magnitude larger than images currently feasible for classification by deep learning methods. To preserve prognostic information contained at high-resolution, WSIs were partitioned into multiple non-overlapping image regions called *tiles* at 10x and 40x resolutions, where each

pixel at 40x represents a physical size of approximately 0·24x0·24 μm^2. Patients without tiles were excluded.

Classification

Five networks were trained on the 634,564 10x tiles and five networks on the 11,591,555 40x tiles from the 1652 Aperio AT2 and NanoZoomer XR WSIs in the training cohort with the patients' distinct outcomes as ground-truth. All networks were DoMore v1 networks, which we designed for classifying supersized heterogeneous images. The DoMore v1 network was built around multiple instance learning and comprised of a MobileNetV2[21] representation network, a Noisy-AND pooling function,[22] and a fully-connected classification network similar to the one used by Kraus et al[22] (figure 1). Because of spatial heterogeneity, labelling a tile with the label of its WSI might be problematic. Instead, the networks were trained on labelled collections of tiles. A collection contained tiles from a single WSI, which label it inherits. Collections of tiles were processed by the representation network before the resulting tile representations were pooled and classified. The entire network was trained end-to-end, i.e. directly from image to patient outcome, and each training iteration used a batch size of 32 collections with 64 tiles each. This many tiles were possible because we utilised a novel gradient approximation technique which substantially reduce memory usage during training (appendix pp 4-6). The Noisy-AND pooling function applied a trained non-linear function on tile representation averages. This enhances robustness against tiles not representing the ground-truth, and together with the large number of tiles, alleviates the issues of spatial heterogeneity. During inference, the network processed all tiles in the WSI.

The networks were trained beyond apparent convergence using TensorFlow 1·10, and a model was selected from each network training using the performance in the tuning cohort with the c-index as metric, resulting in five models for each resolution (protocol pp 11-20). Each of the five models provides a score reflecting the probability of poor outcome, and the

average was defined as the ensemble score. For use in categorical markers, suitable thresholds for the 10x and the 40x ensemble scores were determined by evaluations in the tuning cohort to define the ensemble classifiers (protocol pp 20-22). Furthermore, evaluations in the test cohort indicated that combining 10x and 40x markers might be desirable, and two such markers were defined, one continuous and one categorical. The continuous DoMore-v1-CRC score was defined as the average of the 10x and the 40x ensemble scores. The categorical DoMore-v1-CRC classifier assigned to good prognosis if both ensemble classifiers predicted good outcome, uncertain if the ensemble classifiers predicted differently, and poor prognosis if both predicted poor outcome. In a post-hoc analysis, the continuous DoMore-v1-CRC score was categorised into five risk groups (appendix p 6).

Inception v3, a state-of-the-art convolutional neural network, was trained, tuned, and evaluated with the same study setup as the DoMore v1 network (protocol pp 11-22), and tested as a secondary analysis (protocol p 27). While the DoMore-v1-CRC marker was trained using multiple instance learning, each single tile was labelled with the label of its WSI in training the Inception v3 marker. The image distortion algorithm and network hyperparameters were determined independently of the DoMore v1 network in the discovery phase, resulting in slightly different choices for the Inception v3 network (protocol pp 15-16).

Statistical Analysis

This study conformed to the REMARK guideline[23] and relevant aspects of the guideline proposed by Luo et al[24] (appendix pp 7-8). Primary and secondary analyses were planned in advance of evaluations in the validation cohort and described in the protocol.

The pre-defined primary analysis for each scanner was univariable cancer-specific survival (CSS) analysis of the DoMore-v1-CRC classifier; for simplicity, we first present results for the Aperio AT2 scanner and in a separate paragraph address scanner differences. The classifier was included as the only variable in a Cox model to compute the hazard ratio (HR)

with 95% confidence interval (CI) of patients with uncertain and poor prognosis relative to patients with good prognosis. The proportional hazards assumption was found satisfactory fulfilled using log-log plots (appendix p 26). The Mantel-Cox log-rank test was used to assess whether the classifier predicted CSS.

Both the classifier and the continuous score were evaluated in multivariable Cox models as secondary and post-hoc analyses, including markers available at the time of analysis (patients with at least one missing value were excluded). To calculate classification metrics for 3-year CSS, patients without event and less than 3-year follow-up were excluded and events after 3 years were ignored. Category-free net reclassification improvement (NRI) was computed using the Kaplan-Meier estimates of five-year CSS. Two-sided $p<0{\cdot}05$ was considered statistically significant. The confidence level of CIs is 95%. The bias-corrected and accelerated bootstrap CI were computed for NRIs, c-indices and areas under the curves (AUCs) using 10,000 bootstrap replicates and an acceleration constant estimated using leave-one-out cross-validation. Time to CSS in the validation cohort was calculated from date of randomisation to date of cancer-specific death or loss to follow-up. Survival analyses were carried out in Stata/SE 15·1 (StataCorp, TX).

Role of the funding source

The funders had no role in study design, data collection, data analysis, data interpretation, writing the report, or the decision to submit the paper for publication. The corresponding author had full access to all data and the final responsibility to submit for publication.

Results

The DoMore-v1-CRC classifier was a strong predictor of CSS in the primary analysis of the validation cohort (HR for uncertain *vs* good prognosis, 1·89; CI, 1·14-3·15; HR for poor *vs* good prognosis, 3·84; CI, 2·72-5·43; figure 2A). The classifier remained strong in

multivariable analysis (HR for uncertain *vs* good prognosis, 1·56; CI, 0·92-2·65; HR for poor *vs* good prognosis, 3·04; CI, 2·07-4·47; table 2) adjusting for established prognostic markers significant in univariable analyses; pN stage, pT stage, lymphatic invasion, and venous vascular invasion (appendix p 9).

The sensitivity was 52% (CI, 41%-63%), specificity 78% (CI, 75%-81%), positive predictive value 19% (CI, 14%-25%), negative predictive value 94% (CI, 92%-96%), and correct classification rate 76% (CI, 73%-79%) when comparing 3-year CSS to good prognosis *vs* uncertain and poor prognosis. Compared to good and uncertain prognosis *vs* poor prognosis, the sensitivity was 69% (CI, 58%-78%), specificity 66% (CI, 63%-69%), positive predictive value 17% (CI, 13%-21%), negative predictive value 96% (CI, 94%-97%), and correct classification rate 67% (CI, 63%-69%).

The constituents of the DoMore-v1-CRC classifier, the 10x and the 40x ensemble classifiers, were strong predictors in univariable (appendix p 27) and multivariable analyses (appendix pp 10-11). The ensemble classifiers performed similarly as the best classifiers based on one of the ten individual models that constituted the ensemble models (appendix pp 12 and 28-29). The continuous ensemble scores were also strong predictors in univariable (appendix p 9) and multivariable analyses (appendix pp 13-15). The DoMore-v1-CRC score associated strongly with the patient outcome (appendix p 30), and provided a c-index of 0·674 (CI, 0·624-0·719; appendix p 16) in all validation patients and an AUC of 0·713 (CI, 0·624-0·789; appendix p 31) in patients with distinct outcome. The c-index and AUC of the 10x ensemble score were similar to the ones obtained for the DoMore-v1-CRC score (appendix pp 16 and 31).

The DoMore-v1-CRC classifier was a significant predictor of CSS in stage II (HR for poor *vs* good prognosis, 2·71; CI, 1·25-5·86; figure 2C) and stage III (HR for poor *vs* good prognosis, 4·09; CI, 2·77-6·03; figure 2D), and this was confirmed in multivariable analysis (table 2) and for the continuous score (appendix pp 9 and 13). The categorical marker identified patient

groups with substantially different CSS in stage IIIB and IIIC (appendix p 32), and was also significant in pN stages (figures 2C, E, and F) and pT stages (pT1-3 *vs* pT4; appendix p 33). The category-free NRI of supplementing substage with the DoMore-v1-CRC class for prediction of five-year CSS was 61·6% (CI, 43·5%-79·3%); the event-NRI was 3·2% (CI, -13·2%-20·0%), and the non-event-NRI was 58·3% (CI, 52·7%-63·8%).

The DoMore-v1-CRC classifier correlated with a number of factors such as age, pN stage, pT stage, histological grade, location, tumour sidedness, *BRAF* mutation, and microsatellite instability (table 3). Of special interest is the relation to the histopathological grading into well, moderately, and poorly differentiated tumours. This was further studied in the test cohort where all gradings were centrally reviewed by one highly experienced pathologist (NAS).[18,19] Among 133 tumours characterised as well differentiated, the DoMore-v1-CRC classifier assigned 101 as good prognosis, 18 as uncertain and 14 as poor prognosis (appendix p 17). The moderately differentiated tumours were distributed fairly evenly over the DoMore-v1-CRC classes, while among 292 poorly differentiated tumours, the marker assigned 223 as poor prognosis, 36 as uncertain, and 33 as good prognosis. Thus, the DoMore-v1-CRC class was clearly associated to tumour differentiation. The large proportion of tumours classified as moderately differentiated (e.g. 53% [489 of 920] in the test cohort and 75% [846 of 1122] in the validation cohort) restricts the usefulness of this grading system, but also these patients could be risk stratified by the DoMore-v1-CRC marker (appendix p 34).

Median processing time per patient for the entire classification pipeline, i.e. from scan to predicted patient outcome, was 2·8 minutes (interquartile range, 1·8-3·9) in the validation cohort on a computer with an NVIDIA GeForce RTX 2080 Ti and an Intel Core i7-7700K. Inception v3 provided a marker of CSS with only slightly worse performance than the DoMore-v1-CRC classifier (appendix pp 16 and 35-36).

In the test cohort with slides prepared at a different hospital, the classifier provided similar HRs (appendix p 37) as in the validation cohort (figure 2), supporting that it is robust against inter-laboratory differences in tissue preparation and staining.

When evaluated using another scanner (NanoZoomer XR), the DoMore-v1-CRC score tended towards slightly higher values compared to when evaluated using the Aperio AT2 scanner, resulting in a higher DoMore-v1-CRC class for some patients near the classification thresholds (appendix p 38). However, the scores correlated strongly (Pearson's r=0·956; CI, 0·951-0·961), and the classifier provided similar prognostic information with both scanners (see appendix pp 9, 16, 18-25, and 39-51 for results with NanoZoomer XR). Thus, the classifier was also a strong predictor of CSS in the primary analysis of the validation cohort when evaluated on NanoZoomer XR slide images (HR for uncertain *vs* good prognosis, 2·42; CI, 1·45-4·03; HR for poor *vs* good prognosis, 3·39; CI, 2·36-4·87; appendix p 39).

Discussion

Building on recent developments in machine learning, we have developed a biomarker for automatic prediction of the outcome of a patient resected for early-stage CRC which directly analyse standard H&E stained histological sections. To assay the biomarker, one convolutional neural network first automatically outlines cancerous tissue, and then a second convolutional neural network stratifies the patients into prognostic categories. In the validation, the good and poor prognosis groups included nearly 90% of the patients and differed about 4 times in HR for CSS in univariable analysis and about 3 times in multivariable analysis. The multivariable result indicated that the new biomarker will be a useful supplement to the established markers and improve risk stratification.

Deep learning has already been shown to be suitable for detection and delineation of some tumour types,[25] and various cancer classifications have been reported.[26] Recent studies have

suggested that deep learning could be used to develop markers which potentially utilise basic morphology to predict the outcome of cancer patients, but these findings have not been validated in independent cohorts.[27,28] We have not yet seen independently validated markers for directly predicting the outcome of cancer patients based on histological images.

We derived two markers using the same study setup, but different deep learning techniques. In training the Inception v3 marker, each tile was labelled with the label of its WSI, while the DoMore-v1-CRC marker was developed using multiple instance learning to allow training on tile collections labelled with the label of its WSI. Both markers were strong predictors of CSS, but the DoMore-v1-CRC marker performed slightly better and was the marker pre-selected for independent validation in the QUASAR 2 cohort.

Automatic prognostication procedures reduce human intervention, and has the potential to increase reproducibility of biomarkers. New procedures like the DoMore-v1-CRC markers may initially be performed as services carried out at specialised laboratories with a high degree of standardisation of procedure to avoid disparities in sample handling, including the staining and scanning. Such centralised processing will also facilitate the collection of information on new procedures and enable improvements in the decision support to pathologists and clinicians. As an increasing number of laboratories are becoming digitalised, accompanying decision support systems may include standardisation modules and facilitate a more rapid spread of the automatic procedures. Moreover, supplemented by increased robotisation of wet-lab procedures, the higher analytic throughput will allow decisions based on multiple samples from a tumour. This may reduce the challenge of tumour heterogeneity, which may be a key to improved accuracy of prognosis.

The DoMore-v1-CRC biomarker correlated with several recognised prognostic factors, including the histological grading carried out by a specialised pathologist. The classifier performed better than most other markers in terms of HRs in stage-specific multivariable

analyses, on a par with pN staging. As opposed to the grading system, the classifier had few patients in the intermediate "uncertain" group.

The DoMore-v1-CRC classifier is technically simple to apply and can be delivered at pathology laboratories everywhere. Although training the networks was resource demanding, new patients can be assayed in a few minutes using consumer hardware.

Clinically, the marker will inform discussion with patients with stage II and III CRC on the pros and cons of different adjuvant treatment options. Although the number of drugs used in the adjuvant setting is limited to fluoropyrimidines ± oxaliplatin, recent data demonstrate that three months treatment achieves approximately the same survival outcomes as six months for the majority of stage III patients, while high risk patients (pT4 and pN2) might benefit from prolonged therapy.[29,30] It would be reasonable to hypothesise that stage III patients identified as poor prognosis by the DoMore-v1-CRC classifier could benefit from prolonged combination chemotherapy with oxaliplatin, or even consider experimental therapy combining fluoropyrimidine + oxaliplatin + irinotecan as their high risk of cancer-specific death should positively skew the risk-benefit ratio of more aggressive treatments (figures 2D and F). At the other end, stage III patients with DoMore-v1-CRC good prognosis, the great majority of whom are pN1, have very good survival with single-agent capecitabine (figure 2E), and good prognosis stage II patients have a very high chance of surgical cure, potentially eliminating the need for adjuvant treatment.

We plan to undertake prospective adjuvant trials stratifying patients into different prognostic groups using the DoMore-v1-CRC biomarker and randomising patients into observation, low intensity and high intensity regimes depending on relative risk score. However, the currently available data may also be used by clinicians and patients to make joint and more informed decisions on adjuvant chemotherapy choices, as the proportional reduction in the HRs for recurrence and death from CRC following adjuvant treatment is remarkably consistent at 20%

across most well-designed clinical trials, thus translating into quite different absolute survival improvements for low and high risk subgroups.

Limitation of this study include that the DoMore-v1-CRC marker has not yet been tested prospectively in clinical settings, and although we are planning a clinical trial with randomisation, we at present only know the outcome of thorough retrospective testing. The test and validation indicate good transferability between populations, but there are still challenges related to standardisation, as illustrated by the differences between the tested scanners. Differences between laboratories may also be seen for sample handling procedures, and this is why the introduction into the clinic is suggested to be through services performed at specialised laboratories. A well-known disadvantage of deep learning is its black-box nature. The DoMore-v1-CRC marker is related to histological grading, but the marker is still using small-scale features of the histological images with unknown biological correlates.

In summary, it has been possible to develop a clinically useful prognostic marker using deep learning allied to digital scanning of conventional H&E stained, FFPE tumour tissue sections. The assay has been extensively evaluated in large, independent patient populations, correlates with and outperforms established molecular and morphological prognostic markers, gives consistent results across tumour and nodal stage, and can potentially be used by clinicians to improve decision making over adjuvant treatment choices.

Contributors

OJS, SDR, AK, TSH, KL, FA, DJK, and HED designed the study. HAA, JAN, AN, NAS, IT, RK, MN, and DJK collected the samples and acquired the image data. MP, INF, ED, DNC, AN, NAS, IT, RK, MN, and DJK provided clinical/pathological data and interpretations. OJS, SDR, and JM performed the machine learning. AK performed the statistical analyses. OJS, SDR, AK, TSH, KL, DJK, and HED interpreted the data and analyses. All authors vouch for

the data, analyses, and interpretations. OJS, SDR, AK, TSH, KL, DJK, and HED wrote the first draft of the manuscript, and all authors reviewed, contributed to, and approved the manuscript.

Declaration of interests

OJS, TSH, KL, JM, and HED report filing of a patent application entitled "Histological image analysis" with International Patent Application Number PCT/EP2018/080828. The University of Oxford (to DJK) received educational grants from Roche to support the QUASAR 2 trial and from Merck to support the VICTOR trial. All other authors declare no competing interests.

Acknowledgements

We thank Akershus University Hospital for access to their patient material, National Institute for Health Research for funding support to Marco Novelli through Biomedical Research Centres, Paul Callaghan for animating the appendix video, Marian Seiergren for creating figure 1 and assembling figure 2, the laboratory and technical personnel at the Institute for Cancer Genetics and Informatics for assistance, and the reviewers for valuable suggestions. We also would like to thank the participating centres in the VICTOR and QUASAR 2 trials as well as the staff at Akershus University Hospital, Aker University Hospital and the Gloucestershire hospitals contributing to the Gloucester Colorectal Cancer Study, and last, but not least all participating patients for making this study possible.

References

1. La Thangue NB, Kerr DJ. Predictive biomarkers: a paradigm shift towards personalized cancer medicine. *Nat Rev Clin Oncol* 2011; **8**: 587–96.

2. Van Allen EM, Wagle N, Stojanov P, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med* 2014; **20**: 682–88.

3. Moscow JA, Fojo T, Schilsky RL. The evidence framework for precision cancer medicine. *Nat Rev Clin Oncol* 2018; **15**: 183–92.

4. Karapetis CS, Khambata-Ford S, Jonker DJ, et al. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *N Engl J Med* 2008; **359**: 1757–65.

5. Kerr DJ, Shi Y. Biological markers: Tailoring treatment and trials to prognosis. *Nat Rev Clin Oncol* 2013; **10**: 429–30.

6. Hutchins G, Southward K, Handley K, et al. Value of mismatch repair, KRAS, and BRAF mutations in predicting recurrence and benefits from chemotherapy in colorectal cancer. *J Clin Oncol* 2011; **29**: 1261–70.

7. Salazar R, Roepman P, Capella G, et al. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. *J Clin Oncol* 2011; **29**: 17–24.

8. Gray RG, Quirke P, Handley K, et al. Validation study of a quantitative multigene reverse transcriptase-polymerase chain reaction assay for assessment of recurrence risk in patients with stage II colon cancer. *J Clin Oncol* 2011; **29**: 4611–19.

9. QUASAR Collaborative Group. Comparison of fluorouracil with additional levamisole, higher-dose folinic acid, or both, as adjuvant chemotherapy for colorectal cancer: a randomised trial. *Lancet* 2000; **355**: 1588–96.

10. QUASAR Collaborative Group. Adjuvant chemotherapy versus observation in patients with colorectal cancer: a randomised study. *Lancet* 2007; **370**: 2020–29.

11. Andre T, Boni C, Navarro M, et al. Improved overall survival with oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment in stage II or III colon cancer in the MOSAIC trial. *J Clin Oncol* 2009; **27**: 3109–16.

12. Andre T, de Gramont A, Vernerey D, et al. Adjuvant Fluorouracil, Leucovorin, and Oxaliplatin in Stage II to III Colon Cancer: Updated 10-Year Survival and Outcomes According to BRAF Mutation and Mismatch Repair Status of the MOSAIC Study. *J Clin Oncol* 2015; **33**: 4176–87.

13. Sinicrope FA. DNA mismatch repair and adjuvant chemotherapy in sporadic colon cancer. *Nat Rev Clin Oncol* 2010; **7**: 174–77.

14. Mouradov D, Domingo E, Gibbs P, et al. Survival in stage II/III colorectal cancer is independently predicted by chromosomal and microsatellite instability, but not by specific driver mutations. *Am J Gastroenterol* 2013; **108**: 1785–93.

15. Danielsen HE, Hveem TS, Domingo E, et al. Prognostic markers for colorectal cancer: estimating ploidy and stroma. *Ann Oncol* 2018; **29**: 616–23.

16. Kerr RS, Love S, Segelov E, et al. Adjuvant capecitabine plus bevacizumab versus capecitabine alone in patients with colorectal cancer (QUASAR 2): an open-label, randomised phase 3 trial. *Lancet Oncol* 2016; **17**: 1543–57.

17. Bondi J, Husdal A, Bukholm G, Nesland JM, Bakka A, Bukholm IR. Expression and gene amplification of primary (A, B1, D1, D3, and E) and secondary (C and H) cyclins in colon adenocarcinomas and correlation with patient outcome. *J Clin Pathol* 2005; **58**: 509–14.

18. Petersen VC, Baxter KJ, Love SB, Shepherd NA. Identification of objective pathological prognostic determinants and models of prognosis in Dukes' B colon cancer. *Gut* 2002; **51**: 65–69.

19. Mitchard JR, Love SB, Baxter KJ, Shepherd NA. How important is peritoneal involvement in rectal cancer? A prospective study of 331 cases. *Histopathology* 2010; **57**: 671–79.

20. Midgley RS, McConkey CC, Johnstone EC, et al. Phase III randomized trial assessing rofecoxib in the adjuvant setting of colorectal cancer: final results of the VICTOR trial. *J Clin Oncol* 2010; **28**: 4575–80.

21. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2018: 4510–20.

22. Kraus OZ, Ba JL, Frey BJ. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics* 2016; **32**: i52–i59.

23. Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. *BMC Med* 2012; **10**: 51.

24. Luo W, Phung D, Tran T, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res* 2016; **18**: e323.

25. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA* 2017; **318**: 2199–210.

26. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med* 2018; **24**: 1559–67.

27. Bychkov D, Linder N, Turkki R, et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci Rep* 2018; **8**: 3395.

28. Mobadersany P, Yousefi S, Amgad M, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc Natl Acad Sci U S A* 2018; **115**: E2970–E79.

29. Grothey A, Sobrero AF, Shields AF, et al. Duration of Adjuvant Chemotherapy for Stage III Colon Cancer. *N Engl J Med* 2018; **378**: 1177–88.

30. Iveson TJ, Kerr RS, Saunders MP, et al. 3 versus 6 months of adjuvant oxaliplatin-fluoropyrimidine combination therapy for colorectal cancer (SCOT): an international, randomised, phase 3, non-inferiority trial. *Lancet Oncol* 2018; **19**: 562–78.

Figure Legends

Figure 1: Pipeline of DoMore-v1-CRC classification

Top: A whole-slide image (WSI) is segmented, and the segmented regions tiled at 40x resolution and 10x resolution. For each resolution, the five trained models each produce one score reflecting the probability of poor outcome. The average of those scores is the ensemble score, one for 10x and one for 40x. If the ensemble score is above a certain threshold, the WSI is classified as poor prognosis. The DoMore-v1-CRC class is determined by the agreement between the two ensemble classifications. Bottom: The DoMore v1 network is comprised of a representation network (MobileNetV2[21]), a pooling function (Noisy-AND[22]), and a simple fully-connected classification network. All components of the DoMore v1 network involve trainable parameters, and the entire network is trained end-to-end. All tiles from a WSI are processed by the representation network one by one, resulting in a collection of tile representations. The pooling function reduces the representations into two numbers, which are then processed by the classification network to produce the score outputted by the model.

Figure 2: **Kaplan-Meier analysis of cancer-specific survival by DoMore-v1-CRC class evaluated on Aperio AT2 slide images in the QUASAR 2 validation cohort**

(A) The primary analysis; all patients evaluated with the pre-defined DoMore-v1-CRC classifier. (B) A post-hoc analysis; all patients evaluated with the DoMore-v1-CRC classifier variant with five categories. (C) A secondary analysis; stage II (equivalent to pN0) patients evaluated with the pre-defined DoMore-v1-CRC classifier. (D) A secondary analysis; stage III patients evaluated with the pre-defined DoMore-v1-CRC classifier. (E) A post-hoc analysis; pN1 patients evaluated with the pre-defined DoMore-v1-CRC classifier. (F) A post-hoc analysis; pN2 patients evaluated with the pre-defined DoMore-v1-CRC classifier.

*Manuscript with revisions highlighted

Deep learning for prediction of colorectal cancer outcome: a discovery and validation study

Ole-Johan Skrede, M. Sc.[1,2,*], Sepp De Raedt, Ph. D.[1,2,*], Andreas Kleppe, Ph. D.[1,2], Tarjei S. Hveem, Ph. D.[1], Prof. Knut Liestøl, Ph. D.[1,2], John Maddison, Ph. D.[1], Hanne A. Askautrud, Ph. D.[1], Manohar Pradhan, Ph. D.[1], John Arne Nesheim, M. Sc.[1], Prof. Fritz Albregtsen, M. Sc.[1,2], Prof. Inger Nina Farstad, Ph. D.[3,4], Enric Domingo, Ph. D.[5 6,7], David N. Church, D. Phil.[5,6,7], Prof. Arild Nesbakken, Ph. D.[4,8,9], Prof. Neil A. Shepherd, D. M.[10], Prof. Ian Tomlinson, Ph. D.[1,11], Prof. Rachel Kerr, Ph. D.[75], Prof. Marco Novelli, Ph. D.[1,12], Prof. David J. Kerr, D. Sc.[13], Prof. Håvard E. Danielsen, Ph. D.[1,2,13]**

[1]Institute for Cancer Genetics and Informatics, Oslo University Hospital, Oslo, Norway

[2]Department of Informatics, University of Oslo, Oslo, Norway

[3]Department of Pathology, Division of Laboratory Medicine, Oslo University Hospital, Oslo, Norway

[4]Institute of Clinical Medicine, University of Oslo, Oslo, Norway

[5]~~NIHR~~[5]Department of Oncology, University of Oxford, Oxford, UK

[6]NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, UK

[6]~~Wellcome~~[7]Wellcome Centre for Human Genetics~~, University of Oxford, Oxford, UK~~

[7]~~Department of Oncology~~, University of Oxford, Oxford, UK

[8]Department of Gastrointestinal Surgery, Oslo University Hospital, Oslo, Norway

[9]K.G. Jebsen colorectal cancer research centre, Oslo, Norway

[10]Gloucestershire Cellular Pathology Laboratory, Cheltenham General Hospital, Cheltenham, UK

Style Definition: Normal

Formatted: Not Superscript/ Subscript

Formatted: Superscript

[11]~~Cancer Genetics and Evolution Laboratory, Institute of~~[11]Edinburgh Cancer ~~and Genomic Sciences~~Research Centre, University of ~~Birmingham, Edgbaston, Birmingham, UK~~Edinburgh, Edinburgh, Scotland

[12]Research Department of Pathology, University College London Medical School, London, UK

[13]Nuffield Division of Clinical Laboratory Sciences, University of Oxford, Oxford, UK

*Both authors contributed equally to this work.

**Corresponding author:

Prof Håvard E. Danielsen,

Institute for Cancer Genetics and Informatics,

Oslo University Hospital

Montebello, 0310, Oslo, Norway

Email: hdaniels@labmed.uio.no

Phone: +47 22782320

Words in abstract (not exceed 300): 297

Words in main text (up to 3500): ~~3655~~3889

Number of references (up to 30): 30

Number of figures: 2

47 Number of tables: 3

48

Formatted: Space After: 0 pt

Background: Improved markers of prognosis are needed to stratify patients with early-stage colorectal cancer to refine selection of adjuvant therapy. The aim of the present study was to develop a biomarker of patient outcome after primary colorectal cancer resection by directly analysing scanned conventional haematoxylin and eosin stained sections using deep learning.

Methods: More than 12,000,000 image tiles from 828 patients with distinctly good or poor disease outcome were used to train a total of 10 convolutional neural networks, purpose-built for classifying supersized heterogeneous images. A prognostic biomarker integrating the 10 networks were determined using 1645 patients with non-distinct outcome. The marker was tested on 920 patients with slides prepared in UK, and finally independently validated according to a pre-defined protocol in 1122 patients treated with single-agent capecitabine using slides prepared in Norway. The primary outcome was cancer-specific survival.

Findings: The biomarker provided a hazard ratio for poor *vs* good prognosis of 3·84 (95% confidence interval, 2·72-5·43; p<0·0001) in the primary analysis of the validation cohort, and 3·04 (95% confidence interval, 2·07-4·47; p<0·0001) after adjusting for established prognostic markers significant in univariable analyses of the same cohort; pN stage, pT stage, lymphatic invasion, and venous vascular invasion.

Interpretation: It was possible to develop a clinically useful prognostic marker using deep learning allied to digital scanning of conventional haematoxylin and eosin stained tumour tissue sections. The assay has been extensively evaluated in large, independent patient populations, correlates with and outperforms established molecular and morphological prognostic markers, and gives consistent results across tumour and nodal stage. The biomarker stratified stage II and III patients into sufficiently distinct prognostic groups that these potentially could be used to guide selection of adjuvant treatment by avoiding therapy in very low risk groups and identifying patients who would benefit from more intensive regimes.

Funding: The Research Council of Norway through its IKTPLUSS Lighthouse program (grant number 259204, project name DoMore!).

Research in context

Evidence before this study

Digital image analysis is one of the fields where the recent renaissance of deep learning has achieved the most impressive results. We searched PubMed on June 12, 2019 without language or time restrictions, using the terms “deep learning”, “prediction”, “survival”, “cancer”, and “histology” (full specification of the search criteria is provided in the appendix p 3). We systematically reviewed the 214 search results, and found 18 original research studies which applied deep learning to predict patient outcome or related attributes using histopathology images.

In 16 studies, the patient outcome was indirectly predicted by identifying attributes known to correlate with patient outcome, e.g. stromal fraction, mitotic count, or Gleason pattern. Two studies reported on direct prediction of survival, but neither presented a marker for automatic prediction of patient outcome from scanned whole-slide sections; one required manual annotation to locate interesting tissue regions, and the other classified tissue microarray spots. Perhaps even more importantly, neither of these two studies evaluated their biomarker in independent cohorts; the performance was instead estimated using cross-validation in the same cohort as utilised for training, which can easily lead to overoptimistic estimates.

Added value of the study

We have applied deep learning to develop a biomarker for automatic prediction of cancer-specific survival directly from scanned haematoxylin and eosin stained, formalin-fixed, paraffin-embedded tumour tissue sections. Independent validation demonstrated that the

biomarker improved prediction of cancer-specific survival by stratifying stage II and III colorectal cancer patients into distinct prognostic groups, supplementing established prognostic markers, and outperforming most existing markers in terms of hazard ratios. The marker could potentially be used to improve selection of adjuvant treatment after resection of colorectal cancer by identifying patients at very low risk who may have been cured by surgery alone, as well as patients at high risk who are much more likely to benefit from more intensive regimes.

Implications of all the available evidence

It is possible to utilise deep learning to develop biomarkers for automatic prediction of patient outcome directly from conventional histopathology images. In colorectal cancer, the marker was found to be a clinically useful prognostic marker in analysis of a large series of patients ~~which~~who received consistent, modern cancer treatment.

Introduction

Biomarkers are being used increasingly to match anticancer therapy to specific tumour genotypes, protein, and RNA expression profiles, usually in patients with advanced disease.[1–3] One example of this is selection of *KRAS*-wild-type colorectal cancers (CRCs) for treatment with epidermal growth factor receptor inhibitors.[4] However, in the adjuvant setting for CRC, the primary question is binary, whether to offer treatment at all, and subsequent selection of drugs, dose, and schedule is predominantly driven by stage rather than by companion diagnostics. If it were possible to further refine prognostic models, this could allow a more targeted approach by defining subgroups in which the absolute benefits of adjuvant chemotherapy are minimal, relative to surgery alone, and at the other end of the spectrum, patients who might benefit from prolonged combination chemotherapy because of their poor survival rate.[5–8]

More than two decades of adjuvant trials in patients with early-stage CRC using ~~fluropyrimidines~~fluoropyrimidines, in combination with cytotoxic agents like oxaliplatin, have yielded an improved overall survival of around 3-5% for patients with stage II or IIIA CRC. Many patients are cured by surgery alone, while around 25% will recur despite adjuvant chemotherapy. There is likely to be a chemotherapy-associated death rate of 0~~.~~·5-1%, and 20% of patients will suffer significant side-effects. The risk-benefit ratio is therefore rather marginal, but could potentially be much better if it were possible to define subgroups at higher or lower risk of recurrence and cancer-specific death.[9–12]

Although clinically validated prognostic biomarkers would facilitate adjuvant therapeutic decisions, very few have been sufficiently robustly validated for routine clinical application. A case can be made for assessment of mismatch repair (MMR) status,[13,14] as patients with MMR-deficient tumours tend to have a good prognosis. We have recently reported that measurement of tumour cellular DNA content (ploidy) in combination with stromal fraction

can stratify stage II patients into very good, intermediate, and poor prognostic groups.[15] Interestingly, analysis of driver mutations and RNA signatures has shown them to be individually weak prognostic markers and unable to guide clinical decision making.[8,14]

Deep learning ~~supersedes other machine learning techniques in many applications and is expected to~~refers to the class of machine learning methods that make use of successively more abstract representations of the input data to perform a specific task. These methods use a training set to learn how these representations should be generated in a manner appropriate for the given task. In contrast, traditional machine learning utilises handcrafted features to create representations of the input data that are applied to perform the task. In many applications, deep learning has been demonstrated to provide superior performance compared to other machine learning techniques, and it is a growing expectation that deep learning will transform current medical practice. Especially convolutional neural networks have excelled in many image interpretation tasks, and could therefore be hypothesised to retrieve additional information from ~~pathological~~histopathology images. The aim of the present study was to use deep learning to analyse conventional whole-slide images (WSIs) in order to develop an automatic prognostic biomarker for patients resected for primary CRC. The marker was trained using 828 patients with distinct prognosis from four cohorts, fine-tuned using 1645 other patients from the same four cohorts, and tested on slides prepared at a different laboratory from 920 patients. Finally, the marker was independently validated according to the pre-defined protocol (appendix pp 52-80) on 1122 patients analysed retrospectively from a trial (QUASAR 2) of adjuvant therapy.[16]

Methods

Training and Tuning Cohorts

Four different cohorts were utilised for training and tuning to achieve a broad patient representation and thereby improve the ability to generalise to new cohorts. Three cohorts were consecutive series of stage I, II or III tumours from CRC patients treated at hospitals with both rural and urban catchment areas: (i) 160 patients treated 1988-2000 at Akershus University Hospital, Norway;[17] (ii) 576 patients treated 1993-2003 at Aker University Hospital, Norway;[15] and (iii) 970 patients treated in Gloucester 1988-1996 and included in the Gloucester Colorectal Cancer Study, ~~UK[18]~~UK.[18,19]~~.~~ The fourth cohort were 767 stage II or III CRC patients treated at 151 UK hospitals in 2002-2004 and included in the VICTOR trial (ISRCTN registry number ISRCTN98278138).[20] Our cohorts included only patients with resectable tumour, and a formalin-fixed, paraffin-embedded (FFPE) tumour tissue block available for analysis.

To obtain clear ground-truth, we used as training cohort the 828 patients with so-called distinct outcome, either good or poor. A patient was assigned to the good outcome group if aged less than 85 years at surgery, had more than six years follow-up after surgery, and had no record of recurrence or cancer-specific death. The poor outcome group consisted of those aged less than 85 years at surgery and suffered cancer-specific death between 100 days (inclusive) and 2·5 years (exclusive) after surgery. Patients not satisfying either of these group criteria were defined as having non-distinct outcome, and these 1645 patients were used for tuning. The protocol specifies additional cohort details, and demographics are summarised in table 1.

Test Cohort

The test cohort consisted of 920 patients from the Gloucester Colorectal Cancer Study, ~~UK[18]~~UK.[18,19]~~.~~ WSIs were obtained from different FFPE tumour tissue blocks than those used in the training and tuning cohorts.

Validation Cohort

The validation cohort consisted of 1122 patients from 170 hospitals in seven countries recruited to the QUASAR 2 trial (ISRCTN registry number ISRCTN45133151).[16] Inclusion criteria were age 18 years or older, CRC adenocarcinoma histologically proven to be R0 M0 stage III or high-risk stage II, primary resection 4-10 weeks before randomisation, WHO performance status score 0 or 1, and life expectancy (with comorbidities, but excluding cancer risk) of at least five years. See protocol pp 22-25 for exclusion criteria and other details. All patients received adjuvant therapy, either capecitabine plus bevacizumab or capecitabine alone, with equal disease-free and overall survival in both trial arms.[16]

Sample Preparation

Slides in VICTOR cohort were prepared in Oxford, UK, while the other slides in the training and tuning cohorts were prepared at the Institute for Cancer Genetics and Informatics (ICGI), Norway. Introducing this variation in the development phase was hypothesised to increase the robustness and generalisability of the trained marker. Slides in the test cohort were prepared as a part of the routine histopathological examination in Cheltenham, UK, and the performance in this cohort should thus indicate the prognostic ability when the marker is assayed at a different laboratory using original slides. Slides in the validation cohort were prepared at ICGI. All slides were made by staining a three µm FFPE tissue block section with haematoxylin and eosin (H&E), and a pathologist (MP) ascertained that it contained tumour. WSIs were acquired at the highest resolution available (referred to as 40x magnification by the manufacturers) on two scanners, an Aperio AT2 (Leica Biosystems, Germany) and a NanoZoomer XR (Hamamatsu Photonics, Japan).

Areas with high tumour content were identified using a segmentation network that ~~were~~was trained on a subset of the training and tuning cohorts (protocol pp 6-10). A WSI with the so-called 40x resolution typically contained an order of 100,000x100,000 pixels, multiple orders of magnitude larger than images currently feasible for classification by deep learning

methods. To preserve prognostic information contained at high-resolution, WSIs were partitioned into multiple non-overlapping image regions called *tiles* at 10x and 40x resolutions, where each pixel at 40x represents a physical size of approximately 0·24x0·24 μm^2. Patients without tiles were excluded.

Classification

Five networks were trained on the 634,564 10x tiles and five networks on the 11,591,555 40x tiles from the 1652 Aperio AT2 and NanoZoomer XR WSIs in the training cohort with the patients' distinct outcomes as ground-truth. All networks were DoMore v1 networks, which we designed for classifying supersized heterogeneous images. The DoMore v1 network was built around multiple instance learning and comprised of a MobileNetV2[21] representation network, a Noisy-AND pooling function,[22] and a fully-connected classification network similar to the one used by Kraus et al[22] (figure 1). Because of spatial heterogeneity, labelling a tile with the label of its WSI might be problematic. Instead, the networks were trained on labelled collections of tiles. A collection contained tiles from a single WSI, which label it inherits. Collections of tiles were processed by the representation network before the resulting tile representations were pooled and classified. The entire network was trained end-to-end, i.e. directly from image to patient outcome, and each training iteration used a batch size of 32 collections with 64 tiles each. This many tiles were possible because we utilised a novel gradient approximation technique which substantially reduce memory usage during training (appendix pp 4-6). The Noisy-AND pooling function applied a trained non-linear function on tile representation averages. This ~~enhance~~enhances robustness against tiles not representing the ground-truth, and together with the large number of tiles, alleviates the issues of spatial heterogeneity. During inference, the network processed all tiles in the WSI.

The networks were trained beyond apparent convergence using TensorFlow 1·10, and a model was selected from each network training using the performance in the tuning cohort

with the c-index as metric, resulting in five models for each resolution (protocol pp 11-20). Each of the five models provides a score reflecting the probability of poor outcome, and the average was defined as the ensemble score. For use in categorical markers, suitable thresholds for the 10x and the 40x ensemble scores were determined by evaluations in the tuning cohort to define the ensemble classifiers (protocol pp 20-22). Furthermore, evaluations in the test cohort indicated that combining 10x and 40x markers might be desirable, and two such markers were defined, one continuous and one categorical. The continuous DoMore-v1-CRC score was defined as the average of the 10x and the 40x ensemble scores. The categorical DoMore-v1-CRC classifier assigned to good prognosis if both ensemble classifiers predicted good outcome, uncertain if the ensemble classifiers predicted differently, and poor prognosis if both predicted poor outcome. In a post-hoc analysis, the continuous DoMore-v1-CRC score was categorised into five risk groups (appendix p 6).

Inception v3, a state-of-the-art convolutional neural network, was trained, tuned, and evaluated with the same study setup as the DoMore v1 network (protocol pp 11-22), and tested as a secondary analysis (protocol p 27). While the DoMore-v1-CRC marker was trained using multiple instance learning, each single tile was labelled with the label of its WSI in training the Inception v3 marker. The image distortion algorithm and network hyperparameters were determined independently of the DoMore v1 network in the discovery phase, resulting in slightly different choices for the Inception v3 network (protocol pp 15-16).

Statistical Analysis

This study conformed to the REMARK guideline[23] and relevant aspects of the guideline proposed by Luo et al[24] (appendix pp 7-8). Primary and secondary analyses were planned in advance of evaluations in the validation cohort and described in the protocol.

The pre-defined primary analysis for each scanner was univariable cancer-specific survival (CSS) analysis of the DoMore-v1-CRC classifier; for simplicity, we first present results for

the Aperio AT2 scanner and in a separate paragraph address scanner differences. The classifier was included as the only variable in a Cox model to compute the hazard ratio (HR) with 95% confidence interval (CI) of patients with uncertain and poor prognosis relative to patients with good prognosis. The proportional hazards assumption was found satisfactory fulfilled using log-log plots (appendix p 26). The Mantel-Cox log-rank test was used to assess whether the classifier predicted CSS.

Both the classifier and the continuous score were evaluated in multivariable Cox models as secondary and post-hoc analyses, including markers available at the time of analysis (patients with at least one missing value were excluded). To calculate classification metrics for 3-year CSS, patients without event and less than 3-year follow-up were excluded and events after 3 years were ignored. Category-free net reclassification improvement (NRI) was computed using the Kaplan-Meier estimates of five-year CSS. Two-sided $p<0{\cdot}05$ was considered statistically significant. The confidence level of CIs is 95%. The bias-corrected and accelerated bootstrap CI were computed for NRIs, c-indices and areas under the curves (AUCs) using 10,000 bootstrap replicates and an acceleration constant estimated using leave-one-out cross-validation. Time to CSS in the validation cohort was calculated from date of randomisation to date of cancer-specific death or loss to follow-up. Survival analyses were carried out in Stata/SE 15·1 (StataCorp, TX).

Role of the funding source

The funders had no role in study design, data collection, data analysis, data interpretation, writing the report, or the decision to submit the paper for publication. The corresponding author had full access to all data and the final responsibility to submit for publication.

Results

The DoMore-v1-CRC classifier was a strong predictor of CSS in the primary analysis of the validation cohort (HR for uncertain *vs* good prognosis, 1·89; CI, 1·14-3·15; HR for poor *vs* good prognosis, 3·84; CI, 2·72-5·43; figure 2A). The classifier remained strong in multivariable analysis (HR for uncertain *vs* good prognosis, 1·56; CI, 0·92-2·65; HR for poor *vs* good prognosis, 3·04; CI, 2·07-4·47; table 2) adjusting for established prognostic markers significant in univariable analyses; pN stage, pT stage, lymphatic invasion, and venous vascular invasion (appendix p 9).

The sensitivity was 52% (CI, 41%-63%), specificity 78% (CI, 75%-81%), positive predictive value 19% (CI, 14%-25%), negative predictive value 94% (CI, 92%-96%), and correct classification rate 76% (CI, 73%-79%) when comparing 3-year CSS to good prognosis *vs* uncertain and poor prognosis. Compared to good and uncertain prognosis *vs* poor prognosis, the sensitivity was 69% (CI, 58%-78%), specificity 66% (CI, 63%-69%), positive predictive value 17% (CI, 13%-21%), negative predictive value 96% (CI, 94%-97%), and correct classification rate 67% (CI, 63%-69%).

The constituents of the DoMore-v1-CRC classifier, the 10x and the 40x ensemble classifiers, were strong predictors in univariable (appendix p 27) and multivariable analyses (appendix pp 10-11). The ensemble classifiers performed similarly as the best classifiers based on one of the ten individual models that constituted the ensemble models (appendix pp 12 and 28-29). The continuous ensemble scores were also strong predictors in univariable (appendix p 9) and multivariable analyses (appendix pp 13-15). The DoMore-v1-CRC score associated strongly with the patient outcome (appendix p 30), and provided a c-index of 0·674 (CI, 0·624-0·719; appendix p 16) in all validation patients and an AUC of 0·713 (CI, 0·624-0·789; appendix p 31) in patients with distinct outcome. The c-index and AUC of the 10x ensemble score were similar to the ones obtained for the DoMore-v1-CRC score (appendix pp 16 and 31).

The DoMore-v1-CRC classifier was a significant predictor of CSS in stage II (HR for poor *vs* good prognosis, 2·71; CI, 1·25-5·86; figure 2C) and stage III (HR for poor *vs* good prognosis, 4·09; CI, 2·77-6·03; figure 2D), and this was confirmed in multivariable analysis (table 2) and for the continuous score (appendix pp 9 and 13). The categorical marker identified patient groups with substantially different CSS in stage IIIB and IIIC (appendix p 32), and was also significant in pN stages (figures 2C, E, and F) and pT stages (pT1-3 *vs* pT4; appendix p 33). The category-free NRI of supplementing substage with the DoMore-v1-CRC class for prediction of five-year CSS was 61·6% (CI, 43·5%-79·3%); the event-NRI was 3·2% (CI, -13·2%-20·0%), and the non-event-NRI was 58·3% (CI, 52·7%-63·8%).

The DoMore-v1-CRC classifier correlated with a number of factors such as age, pN stage, pT stage, histological grade, location, tumour sidedness, *BRAF* mutation, and microsatellite instability (table 3). Of special interest is the relation to the ~~pathological~~histopathological grading into well, moderately, and poorly differentiated tumours. This was further studied in the test cohort where all gradings were centrally reviewed by one highly experienced pathologist (NAS).[18,19] Among 133 tumours characterised as well differentiated, the DoMore-v1-CRC classifier assigned 101 as good prognosis, 18 as uncertain and 14 as poor prognosis (appendix p 17). The moderately differentiated tumours were distributed fairly evenly over the DoMore-v1-CRC classes, while among 292 poorly differentiated tumours, the marker assigned 223 as poor prognosis, 36 as uncertain, and 33 as good prognosis. Thus, the DoMore-v1-CRC class was clearly associated to tumour differentiation. The large proportion of tumours classified as moderately differentiated (e.g. 53% [489 of 920] in the test cohort and 75% [846 of 1122] in the validation cohort) restricts the usefulness of this grading system, but also these patients could be risk stratified by the DoMore-v1-CRC marker (appendix p 34).

Median processing time per patient for the entire classification pipeline, i.e. from scan to predicted patient outcome, was 2~~.~~·8 minutes (interquartile range, 1~~.~~·8-3~~.~~·9) in the validation cohort on a computer with an NVIDIA GeForce RTX 2080 Ti and an Intel Core i7-7700K. ~~Inception v3, a state-of-the-art convolutional neural network, was trained, tuned, and evaluated with the same study setup as the DoMore v1 network (protocol pp 11-22), and~~ Inception v3 provided a marker of CSS with only slightly worse performance than the DoMore-v1-CRC classifier (appendix pp 16 and 35-36).

In the test cohort with slides prepared at a different hospital, the classifier provided similar HRs (appendix p 37) as in the validation cohort (figure 2), supporting that it is robust against inter-laboratory differences in tissue preparation and staining.

When evaluated using another scanner (NanoZoomer XR), the DoMore-v1-CRC score tended towards slightly higher values compared to when evaluated using the Aperio AT2 scanner, resulting in a higher DoMore-v1-CRC class for some patients near the classification thresholds (appendix p 38). However, the scores correlated strongly (Pearson's r=0·956; CI, 0·951-0·961), and the classifier provided similar prognostic information with both scanners (see appendix pp 9, 16, 18-25, and 39-51 for results with NanoZoomer XR). Thus, the classifier was also a strong predictor of CSS in the primary analysis of the validation cohort when evaluated on NanoZoomer XR slide images (HR for uncertain *vs* good prognosis, 2·42; CI, 1·45-4·03; HR for poor *vs* good prognosis, 3·39; CI, 2·36-4·87; appendix p 39).

Discussion

Building on recent developments in machine learning, we have developed a biomarker for automatic prediction of the outcome of a patient resected for early-stage CRC which directly analyse standard H&E stained histological sections. To assay the biomarker, one convolutional neural network first automatically outlines cancerous tissue, and then a second

convolutional neural network stratifies the patients into prognostic categories. In the validation, the good and poor prognosis groups included nearly 90% of the patients and differed about 4 times in HR for CSS in univariable analysis and about 3 times in multivariable analysis. The multivariable result indicated that the new biomarker will be a useful supplement to the established markers and improve risk stratification.

Deep learning has already been shown to be suitable for detection and delineation of some tumour types,[25] and various cancer classifications have been reported.[26] Recent studies have suggested that deep learning could be used to develop markers which potentially utilise basic morphology to predict the outcome of cancer patients, but these findings have not been validated in independent cohorts.[27,28] We have not yet seen independently validated markers for directly predicting the outcome of cancer patients based on histological images.

We derived two markers using the same study setup, but different deep learning techniques. In training the Inception v3 marker, each tile was labelled with the label of its WSI, while the DoMore-v1-CRC marker was developed using multiple instance learning to allow training on tile collections labelled with the label of its WSI. Both markers were strong predictors of CSS, but the DoMore-v1-CRC marker performed slightly better and was the marker pre-selected for independent validation in the QUASAR 2 cohort.

Automatic prognostication procedures reduce human intervention, and has the potential to increase reproducibility of biomarkers. New procedures like the DoMore-v1-CRC markers may initially be performed as services carried out at specialised laboratories with a high degree of standardisation of procedure to avoid disparities in sample handling, including ~~in~~ the staining and scanning. Such centralised processing will also facilitate the collection of information on new procedures and enable improvements in the decision support to pathologists and clinicians. As an increasing number of laboratories are becoming digitalised, accompanying decision support systems may include standardisation modules and facilitate a

more rapid spread of the automatic procedures. Moreover, supplemented by increased robotisation of wet-lab procedures, the higher analytic throughput will allow decisions based on multiple samples from a tumour. This may reduce the challenge of tumour heterogeneity, which may be a key to improved accuracy of prognosis.

The DoMore-v1-CRC biomarker correlated with several recognised prognostic factors, including the histological grading carried out by a specialised pathologist. The classifier performed better than most other markers in terms of HRs in stage-specific multivariable analyses, on a par with pN staging. As opposed to the grading system, the classifier had few patients in the intermediate "uncertain" group.

The DoMore-v1-CRC classifier is technically simple to apply and can be delivered at pathology laboratories everywhere. Although training the networks was resource demanding, new patients can be assayed in a few minutes using consumer hardware.

Clinically, the marker will inform discussion with patients with stage II and III CRC on the pros and cons of different adjuvant treatment options. Although the number of drugs used in the adjuvant setting is limited to fluoropyrimidines ± oxaliplatin, recent data demonstrate that three months treatment achieves approximately the same survival outcomes as six months for the majority of stage III patients, while high risk patients (pT4 and pN2) might benefit from prolonged therapy.[29,30] It would be reasonable to hypothesise that stage III patients identified as poor prognosis by the DoMore-v1-CRC classifier could benefit from prolonged combination chemotherapy with oxaliplatin, or even consider experimental therapy combining ~~fluropyrimidine~~fluoropyrimidine + oxaliplatin + irinotecan as their high risk of cancer-specific death should positively skew the risk-benefit ratio of more aggressive treatments (figures 2D and F). At the other end, stage III patients with DoMore-v1-CRC good prognosis, the great majority of whom are pN1, have ~~excellent~~very good survival with single-

agent capecitabine (figure 2E), and good prognosis stage II patients have a very high chance of surgical cure, potentially eliminating the need for adjuvant treatment.

We plan to undertake prospective adjuvant trials stratifying patients into different prognostic groups using the DoMore-v1-CRC biomarker and randomising patients into observation, low intensity and high intensity regimes depending on relative risk score. However, the currently available data may also be used by clinicians and patients to make joint and more informed decisions on adjuvant chemotherapy choices, as the proportional reduction in the HRs for recurrence and death from CRC following adjuvant treatment is remarkably consistent at 20% across most well-designed clinical trials, thus translating into quite different absolute survival improvements for low and high risk subgroups.

Limitation of this study include that the DoMore-v1-CRC marker has not yet been tested prospectively in clinical settings, and although we are planning a clinical trial with randomisation, we at present only know the outcome of thorough retrospective testing. The test and validation indicate good transferability between populations, but there are still challenges related to standardisation, as illustrated by the differences between the tested scanners. Differences between laboratories may also be seen for sample handling procedures, and this is why the introduction into the clinic is suggested to be through services performed at specialised laboratories. A well-known disadvantage of deep learning is its black-box nature. The DoMore-v1-CRC marker is related to histological grading, but the marker is still using small-scale features of the histological images with unknown biological correlates.

In summary, it has been possible to develop a clinically useful prognostic marker using deep learning allied to digital scanning of conventional H&E stained, FFPE tumour tissue sections. The assay has been extensively evaluated in large, independent patient populations, correlates with and outperforms established molecular and morphological prognostic markers, gives

consistent results across tumour and nodal stage, and can potentially be used by clinicians to improve decision making over adjuvant treatment choices.

Contributors

OJS, SDR, AK, TSH, KL, FA, DJK, and HED designed the study. HAA, JAN, AN, NAS, IT, RK, MN, and DJK collected the samples and acquired the image data. MP, INF, ED, DNC, AN, NAS, IT, RK, MN, and DJK provided clinical/pathological data and interpretations. OJS, SDR, and JM performed the machine learning. AK performed the statistical analyses. OJS, SDR, AK, TSH, KL, DJK, and HED interpreted the data and analyses. All authors vouch for the data, analyses, and interpretations. OJS, SDR, AK, TSH, KL, DJK, and HED wrote the first draft of the manuscript, and all authors reviewed, contributed to, and approved the manuscript.

Declaration of interests

OJS, TSH, KL, JM, and HED report filing of a patent application entitled "Histological image analysis" with International Patent Application Number PCT/EP2018/080828. The University of Oxford (to DJK) received educational grants from Roche to support the QUASAR 2 trial and from Merck to support the VICTOR trial. All other authors declare no competing interests.

Acknowledgements

We thank Akershus University Hospital for access to their patient material, National Institute for Health Research for funding support to Marco Novelli through Biomedical Research Centres, Paul Callaghan for animating the appendix video, Marian Seiergren for creating figure 1 and assembling figure 2, the laboratory and technical personnel at the Institute for

457 Cancer Genetics and Informatics for assistance, and the reviewers for valuable suggestions.
458 We also would like to thank the participating centres in the VICTOR and QUASAR 2 trials as
459 well as the staff at Akershus University Hospital, Aker University Hospital and the
460 Gloucestershire hospitals contributing to the Gloucester Colorectal Cancer Study, and last, but
461 not least all participating patients for making this study possible.
462

References

1. La Thangue NB, Kerr DJ. Predictive biomarkers: a paradigm shift towards personalized cancer medicine. *Nat Rev Clin Oncol* 2011; **8**: 587–96.

2. Van Allen EM, Wagle N, Stojanov P, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med* 2014; **20**: 682–88.

3. Moscow JA, Fojo T, Schilsky RL. The evidence framework for precision cancer medicine. *Nat Rev Clin Oncol* 2018; **15**: 183–92.

4. Karapetis CS, Khambata-Ford S, Jonker DJ, et al. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *N Engl J Med* 2008; **359**: 1757–65.

5. Kerr DJ, Shi Y. Biological markers: Tailoring treatment and trials to prognosis. *Nat Rev Clin Oncol* 2013; **10**: 429–30.

6. Hutchins G, Southward K, Handley K, et al. Value of mismatch repair, KRAS, and BRAF mutations in predicting recurrence and benefits from chemotherapy in colorectal cancer. *J Clin Oncol* 2011; **29**: 1261–70.

7. Salazar R, Roepman P, Capella G, et al. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. *J Clin Oncol* 2011; **29**: 17–24.

8. Gray RG, Quirke P, Handley K, et al. Validation study of a quantitative multigene reverse transcriptase-polymerase chain reaction assay for assessment of recurrence risk in patients with stage II colon cancer. *J Clin Oncol* 2011; **29**: 4611–19.

9. QUASAR Collaborative Group. Comparison of fluorouracil with additional levamisole, higher-dose folinic acid, or both, as adjuvant chemotherapy for colorectal cancer: a randomised trial. *Lancet* 2000; **355**: 1588–96.

10. QUASAR Collaborative Group. Adjuvant chemotherapy versus observation in patients with colorectal cancer: a randomised study. *Lancet* 2007; **370**: 2020–29.

11. Andre T, Boni C, Navarro M, et al. Improved overall survival with oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment in stage II or III colon cancer in the MOSAIC trial. *J Clin Oncol* 2009; **27**: 3109–16.

12. Andre T, de Gramont A, Vernerey D, et al. Adjuvant Fluorouracil, Leucovorin, and Oxaliplatin in Stage II to III Colon Cancer: Updated 10-Year Survival and Outcomes According to BRAF Mutation and Mismatch Repair Status of the MOSAIC Study. *J Clin Oncol* 2015; **33**: 4176–87.

13. Sinicrope FA. DNA mismatch repair and adjuvant chemotherapy in sporadic colon cancer. *Nat Rev Clin Oncol* 2010; **7**: 174–77.

14. Mouradov D, Domingo E, Gibbs P, et al. Survival in stage II/III colorectal cancer is independently predicted by chromosomal and microsatellite instability, but not by specific driver mutations. *Am J Gastroenterol* 2013; **108**: 1785–93.

15. Danielsen HE, Hveem TS, Domingo E, et al. Prognostic markers for colorectal cancer: estimating ploidy and stroma. *Ann Oncol* 2018; **29**: 616–23.

16. Kerr RS, Love S, Segelov E, et al. Adjuvant capecitabine plus bevacizumab versus capecitabine alone in patients with colorectal cancer (QUASAR 2): an open-label, randomised phase 3 trial. *Lancet Oncol* 2016; **17**: 1543–57.

17. Bondi J, Husdal A, Bukholm G, Nesland JM, Bakka A, Bukholm IR. Expression and gene amplification of primary (A, B1, D1, D3, and E) and secondary (C and H) cyclins in colon adenocarcinomas and correlation with patient outcome. *J Clin Pathol* 2005; **58**: 509–14.

18. Petersen VC, Baxter KJ, Love SB, Shepherd NA. Identification of objective pathological prognostic determinants and models of prognosis in Dukes' B colon cancer. *Gut* 2002; **51**: 65–69.

19. Mitchard JR, Love SB, Baxter KJ, Shepherd NA. How important is peritoneal involvement in rectal cancer? A prospective study of 331 cases. *Histopathology* 2010; **57**: 671–79.

20. Midgley RS, McConkey CC, Johnstone EC, et al. Phase III randomized trial assessing rofecoxib in the adjuvant setting of colorectal cancer: final results of the VICTOR trial. *J Clin Oncol* 2010; **28**: 4575–80.

21. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2018: 4510–20.

22. Kraus OZ, Ba JL, Frey BJ. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics* 2016; **32**: i52–i59.

23. Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. *BMC Med* 2012; **10**: 51.

24. Luo W, Phung D, Tran T, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *J Med Internet Res* 2016; **18**: e323.

25. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA* 2017; **318**: 2199–210.

26. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med* 2018; **24**: 1559–67.

27. Bychkov D, Linder N, Turkki R, et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci Rep* 2018; **8**: 3395.

28. Mobadersany P, Yousefi S, Amgad M, et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc Natl Acad Sci U S A* 2018; **115**: E2970–E79.

29. Grothey A, Sobrero AF, Shields AF, et al. Duration of Adjuvant Chemotherapy for Stage III Colon Cancer. *N Engl J Med* 2018; **378**: 1177–88.

30. Iveson TJ, Kerr RS, Saunders MP, et al. 3 versus 6 months of adjuvant oxaliplatin-fluoropyrimidine combination therapy for colorectal cancer (SCOT): an international, randomised, phase 3, non-inferiority trial. *Lancet Oncol* 2018; **19**: 562–78.

Figure Legends

Figure 1: **Pipeline of DoMore-v1-CRC classification**

Top: A whole-slide image (WSI) is segmented, and the segmented regions tiled at 40x resolution and 10x resolution. For each resolution, the five trained models each produce one score reflecting the probability of poor outcome. The average of those scores is the ensemble score, one for 10x and one for 40x. If the ensemble score is above a certain threshold, the WSI is classified as poor prognosis. The DoMore-v1-CRC class is determined by the agreement between the two ensemble classifications. Bottom: The DoMore v1 network is comprised of a representation network (MobileNetV2[21]), a pooling function (Noisy-AND[22]), and a simple fully-connected classification network. All components of the DoMore v1 network involve trainable parameters, and the entire network is trained end-to-end. All tiles from a WSI are processed by the representation network one by one, resulting in a collection of tile representations. The pooling function reduces the representations into two numbers, which are then processed by the classification network to produce the score outputted by the model.

Figure 2: **Kaplan-Meier analysis of cancer-specific survival by DoMore-v1-CRC class evaluated on Aperio AT2 slide images in the QUASAR 2 validation cohort**

(A) The primary analysis; all patients evaluated with the pre-defined DoMore-v1-CRC classifier. (B) A post-hoc analysis; all patients evaluated with the DoMore-v1-CRC classifier variant with five categories. (C) A secondary analysis; stage II (equivalent to pN0) patients evaluated with the pre-defined DoMore-v1-CRC classifier. (D) A secondary analysis; stage III patients evaluated with the pre-defined DoMore-v1-CRC classifier. (E) A post-hoc analysis; pN1 patients evaluated with the pre-defined DoMore-v1-CRC classifier. (F) A post-hoc analysis; pN2 patients evaluated with the pre-defined DoMore-v1-CRC classifier.

Responses to the comments

Many thanks for assessing our revised manuscript and providing useful comments. Please find the responses below in bold. All line numbers are referring to the second revision of the manuscript draft.

Editorial points:

The following points list items that must be included before a manuscript can be considered further. Addressing them at this stage reduces the risk of errors and delays later.

Journals differ in requirements for revisions, so please read the requests below carefully and consult me or http://www.thelancet.com/lancet/information-for-authors for further details or clarification if needed.

Additional tips on artwork are available at
http://www.thelancet.com/pb/assets/raw/Lancet/authors/artwork-guidelines.pdf

Response: In conjunction with submitting the first revision of the manuscript draft, we emailed the artworks (figures 1 and 2 as .ai files) directly to the handling editor. The figures have not been updated since. We would happily provide the figures again upon request.

If your manuscript is a RCT, Formatting guidelines are available at
http://www.thelancet.com/pb/assets/raw/Lancet/authors/Rctguidelines.pdf

Response: Not applicable because our study is not an RCT.

Please note that not every point below will be relevant to your manuscript.

1. Please indicate after each of the reviewers' points the text changes which have been made (if any) and the line number on the revised manuscript at which your change can be found. [Line numbers can be added to your word document using the 'page layout' tab. Please select continuous numbers.]

Response: In the responses to the editorial points and reviewers' comments, the text changes have been indicated by the line numbers on the second revision of the manuscript draft.

2. When interpreting editorial points made by reviewers, please remember we will edit the final manuscript if accepted.

Response: We understand that the manuscript will be edited if accepted.

3. Please indicate any authors who are full professors.

Response: The following authors are full professors: Knut Liestøl, Fritz Albregtsen, Inger Nina Farstad, Arild Nesbakken, Neil A. Shepherd, Ian Tomlinson, Rachel Kerr, Marco Novelli, David J. Kerr, and Håvard E. Danielsen. The professors are indicated by "Prof." in the author list (see lines 4-10).

4. Please list the highest degree for each author (one degree only, please).

Response: Only the single highest degree for each author has been listed in the author list (see lines 4-10).

5. Please check that all author name spellings and affiliations are correct.

Response: We have verified and corrected all author name spellings and affiliations (see lines 4-28).

6. For randomised trials please follow the CONSORT reporting guidelines (http://www.consort-statement.org) and CONSORT for abstracts (http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(07)61835-2/fulltext), and include a CONSORT checklist with your resubmission.

Response: Not applicable because our study is not an RCT.

7. Please ensure that the title of the paper is non-declamatory (ie, it describes the aim of study rather than the findings) and that it includes a description of the study type (eg, a randomised controlled trial).

Response: The manuscript title is non-declamatory and includes a description of the study type.

8. Please limit the summary to pre-defined primary endpoints and safety endpoints.

Response: The summary is limited to the pre-defined primary endpoint. Safety endpoints are not applicable to this retrospective cohort study.

9. For RCTs, please state the trial registration number.

Response: Our study is not an RCT, but two of the cohorts which we analysed retrospectively are from RCTs. Their trial registration numbers are specified in lines 165-166 and 184.

10. At the end of the methods section please state the role of the funder in: data collection, analysis, interpretation, writing of the manuscript and the decision to submit. Please also state which author(s) had access to all the data, and which author(s) were responsible for the decision to submit the manuscript etc.

Response: The role of the funder is stated at the end of the methods section (see lines 276-279).

11. Please explain any deviations from the protocol.

Response: There was no deviation from the protocol.

12. Please report all outcomes specified in the protocol.

Response: All outcomes specified in the protocol are reported in the manuscript.

13. If any exploratory outcomes are reported that were not pre-specified, please make it clear that these analyses were post-hoc.

Response: All reported outcomes were pre-specified in the study protocol, and all exploratory analyses are described as post-hoc analyses in the manuscript.

14. Please use rINNs for drug names. For genes and proteins, authors can use their preferred terminology so long as it is in current use by the community, but should provide the preferred human name from Uniprot (http://www.uniprot.org/uniprot/) for proteins and HUGO (http://www.genenames.org) for genes at first use to assist non-specialists.

Response: Not applicable for our manuscript.

15. For drug studies, please ensure that details of doses, route of delivery, and schedule are included.

Response: Not applicable for our study.

16. For the main outcome measures, please include a result for each group, plus a point estimate (eg, RR, HR) with a measure of precision (eg, 95% CI) for the absolute difference between groups, in both the Summary and the main Results section of the paper.

Response: We have included hazard ratios with 95% confidence intervals in both the Summary (see lines 57-59) and the main Results section (see lines 283-286, 307-308, and 344-345).

17. p-values should be exact, but no longer than 4 decimal places (eg p<0.0001). Two decimals are acceptable in tables for non-significant p-values

Response: Exact p-values are provided with two significant digits, but no longer than 4 decimal places.

18. Please provide absolute numbers to accompany all percentages. Percentages should be rounded to whole numbers unless the study population is very large (>10 000 individuals).

Response: Percentages are rounded to whole numbers and reported with the absolute numbers they were computed from.

19. Please give 95% confidence intervals for hazard ratios/odds ratios.

Response: 95% confidence intervals for hazard ratios have been provided.

20. For means, please provide standard deviation (or error, as appropriate).

Response: Not applicable because we have not reported any means.

21. Please provide interquartile ranges for medians.

Response: Interquartile ranges have been provided for medians (see line 330, and tables 1 and 3).

22. Please provide numbers at risk for Kaplan-Meier plots and ensure that plots include a measure of effect (eg, log-rank p); estimates should be reported with 95% CIs.

Response: Numbers at risk, log-rank p, and hazard ratios with 95% confidence intervals have been provided for Kaplan-Meier plots.

23. Please ensure that the Discussion contains a section on limitations of the study.

Response: The Discussion contains a section on limitations of the study in lines 411-420.

24. Please provide the text, tables, and figures in an editable format. See link above this list for details of acceptable formats for figure files.

Response: The text and tables are in Word format, and the figures are in the editable .ai format.

25. Our production system is not compatible with Endnotes. Please convert to normal text.

Response: The manuscript was converted to normal text without Endnote field codes before it was submitted.

26. If accepted, only 5-6 non-text items (figures, tables, or panels) can be accommodated in the print edition; additional material can be provided in a web appendix. Please indicate which items can go in a web appendix.

Response: The manuscript contains 2 figures and 3 tables, which can be accommodated in the print edition.

27. Please provide a research in context panel with 3 parts: Evidence before this study (which includes a description of how you searched for evidence and how you assessed the quality of that evidence); Added value of the study; and Implications of all the available evidence.

Response: Text for a research in context panel is provided in lines 73-108.

28. At the end of the manuscript, please summarise the contribution of each author to the work.

Response: The contribution of each author has been summarised at the end of the manuscript (see lines 428-436).

29. At the end of the manuscript please summarise the declaration of interests for each author.

Response: The declaration of interests for each author has been summarised at the end of the manuscript (see lines 438-443).

30. If you have not yet done so, please return all signed authorship statements and conflict of interest forms. We also require signed statements from any named person in the acknowledgements saying that they agree to be acknowledged.

Response: The signed authorship statements and conflict of interest forms for all except Prof. Rachel Kerr were uploaded with the first revision of the manuscript draft. The signed authorship statement and conflict of interest form from Prof. Rachel Kerr was emailed directly to the handling editor on 23rd of October and to editorial@lancet.com on 24th of October, and is also uploaded with the second revision of the manuscript draft. Consent forms from the two acknowledged persons were uploaded with the first revision of the manuscript draft.

31. For any personal communication, please provide a letter showing that the person agrees to their name being used.

Response: Not applicable for our manuscript.

32. As corresponding author, please confirm that all authors have seen and approved of the final text.

Response: As the corresponding author, I confirm that all authors have seen and approved the final text.

33. If your author line includes a study group, collaborators' names and affiliations may be listed at the end of the paper or in the appendix. Additionally, if you wish the names of collaborators within a study group to appear on PubMed, please upload with your revision a list of names of all study group members presented as a two-column table in Word. First and middle names or initials should be placed in the first column, and surnames in the second column. Names should be ordered as you wish them to appear on PubMed. The table will not be included in the paper itself - it's simply used to make sure that PubMed adds the names correctly.

Response: Not applicable for our manuscript.

34. Please note our guideline length for research articles is 3500 words and 30 references. For RCTs, the text can be expanded to 4500 words.

Response: We have attempted to present the findings concisely and precisely, also when revising the manuscript according to the reviewers' comments. The second revision of the manuscript draft contains 3889 words and 30 references.

35. From July 1, 2018, all submitted reports of clinical trials must contain a data sharing statement, to be included at the end of the manuscript or in an appendix (please provide as a separate pdf). Data sharing statements must indicate:

*Whether data collected for the study, including individual participant data and a data dictionary defining each field in the set, will be made available to others;

*What data will be made available (deidentified participant data, participant data with identifiers, data dictionary, or other specified data set);

*Whether additional, related documents will be available (eg, study protocol, statistical analysis plan, informed consent form);

*When these data will be available (beginning and end date, or "with publication", as applicable);

*Where the data will be made available (including complete URLs or email addresses if relevant);

*By what access criteria data will be shared (including with whom, for what types of analyses, by what mechanism - eg, with or without investigator support, after approval of a proposal, with a signed data access agreement - or any additional restrictions).

Clinical trials that begin enrolling participants on or after Jan 1, 2019, must include a data sharing plan in the trial's registration. If the data sharing plan changes after registration, this should be reflected in the statement submitted and published, and updated in the registry record. For reports of research other than clinical trials, data sharing statements are encouraged but not required. Mendeley Data (https://data.mendeley.com) is a secure online repository for research data, permitting archiving of any file type and assigning a permanent and unique digital object identifier (DOI) so that the files can be easily referenced. If authors wish to share their supporting data, and have not already made alternative arrangements, a Mendeley DOI can be referred to in the data sharing statement.

Response: Not applicable for our manuscript.

Reviewer #1: Dear Editors and Authors,

Thank you for the opportunity to review this revised article. I have re-read the article and author comments to reviewers. The authors have made an intelligent attempt to address reviewer concerns.

I only have a few minor comments.

Comments:

1) It is recommended the authors define 'deep learning' in lay terms and perhaps contrast this with traditional machine learning in a few words for readers in the introduction section.

Response: Thank you for helping us improve the presentation. We have revised lines 138-147 to introduce deep learning in layman's terms, separating it from traditional machine learning.

2) The authors state "At the other end, stage III patients with DoMore-v1-CRC good prognosis, the great majority of whom are pN1, have very excellent survival with single - agent c 440 apecitabine (figure 2E),". In the previous draft this was stated as 'very good' - now changed to 'excellent'. If the data has changed from the previous draft, then perhaps this should be highlighted here. If not then perhaps the language should be toned down a bit and reflect what was presented in the original draft.

Response: The data has not changed. Line 400 has been revised according to your suggestion, i.e. the language has been toned down by using the original statement "very good" instead of "excellent".

3) I could not find the figures in this draft and therefore could not review them. Please ensure these are submitted to the editors.

Response: We provided the figures in the editable format .ai since an editable format was requested in the 24th editorial point. Since the submission system did not allow us to upload .ai files, the figure files were instead emailed directly to the handling editor. The content of the figures is identical in the original submission and in both revisions.

Reviewer #2: The authors addressed most of the reviewers' concerns. Still, there are a few points left that need to be addressed.

Major:

As requested, the authors added more extensive results for the Inception-V3 approach which is used for comparison. However, the approach is hardly mentioned in the manuscript itself as most of the results are in the appendix. The Inception-V3 model should be briefly introduced in the methods part and a few comments could be added in the discussion. In particular, the authors could briefly discuss the advantage of using the DoMore approach over the Inception-V3 approach for clinical practice. This might not be obvious to the reader as the performance of the two approaches is close and, potentially, both could serve as a useful marker. Of course, the key contribution of this paper is to demonstrate that a deep learning-based marker could be useful for prognosis. For this contribution, the DoMore approach does not necessarily have to be better than other options (e.g. Inception-V3) for clinical use. In any case, this should be cleared up in the discussion.

Response: In accordance to your suggestions, we have added lines 245-251 to introduce briefly the Inception v3 marker and lines 363-368 to discuss briefly the conceptual and practical differences between the two markers, noting in particular that both markers performed well, but that the DoMore-v1-CRC marker appears better and was the marker pre-selected for independent validation in the QUASAR 2 cohort (see protocol pp 25-26).

Minor:

Referring to:

Protocol, p.15-16, description of Inception V3 training: The authors do not appear to use the data augmentation techniques as used for the DoMore architecture. In particular, there appears to be no random cropping and flipping/rotation. This would make the comparison not very meaningful as Inception-V3 might perform better with the same data augmentation scheme. This should be clarified.

Response: While the study setup was identical for the two classification setups, the preprocessing and hyperparameters were adapted to the specific network (DoMore v1 or Inception v3) in the discovery phase. This would ideally provide a fairer comparison between the potential of each network because the preprocessing and hyperparameters are then not adapted to one of the networks and possibly inappropriate for the other.

Reviewer comment: From my understanding, the authors indicate that using the additional data augmentation used for DoMore (namely random cropping and flipping/rotation) does not increase performance for Inception-V3. Since these are standard augmentation techniques I am very surprised that they are not helpful for Inception-V3. The authors should state that they performed individual hyperparameter tuning for each approach (DoMore and Inception-V3) beforehand which resulted in the selection of the different data augmentation techniques.

Response: Thank you for helping us improve our presentation. We now specify in the revised lines 245-246 and 249-251 that while the study setup was identical for the two approaches (DoMore v1 and Inception v3), the image distortion algorithm and hyperparameters were independently determined in the discovery phase. The impact of applying random cropping and flipping/rotation in training may be much less in our study setup than in many other setups because of our vast amount of training images (i.e. the number of tiles used for training); even when trained beyond apparent convergence, the training of Inception v3 only ran for 3.78 epochs for the 10x networks and 0.41 epochs for the 40x networks. Combined with the applied colour distortion, which was arguably more comprehensive for the Inception v3 training than for the DoMore v1 training (see protocol pp 14-15), it should be highly unlikely that the networks learn to associate many features unique to the training images with the patient outcome. The markers' consistently strong performances in the independent cohorts also indicate that overfitting has not severely reduced the generalised performance of the markers.

Referring to:

* The code and dataset (or at least the code + model weights) should be posted online for replication of the study and to facilitate future research in this domain.

Response: Our goal is to improve the management of many cancer patients. Since commercialisation may be necessary to facilitate widespread adaption in routine medical practice, the Research Council of Norway encouraged projects in the IKTPLUSS Lighthouse program to commercialise products supported by the grant. We currently evaluate such possibilities, and are therefore at present not able to provide code and model weights to the public community. We have however endeavoured to describe all methods with full details, which should enable other researchers to apply the same principles in their own studies.

Reviewer comment: It is understandable that further commercialization could happen. However, this prevents the reproduction of the paper's results. Can the authors comment on the option of making the data publicly available in some way? In my point of view, this would benefit the scientific community much more than a code/model weights release.

Response: Thank you for your understanding. We agree that well-curated datasets could benefit the scientific community and will attempt to solve any ethical, legal and practical issues concerning publicising the raw data that were analysed in this study. This involves a series of institutions in multiple countries, and the raw data cannot be made publically available until all parties have agreed and verified that such distribution does not violate any obligations they may have to the patients, institutions or governments.

Table 1

Table 1: **Patient characteristics in the training, tuning, test and validation cohorts**

	Group	Training cohort (N=828)	Tuning cohort (N=1645)	Test cohort (N=920)	Validation cohort (N=1122)
Age, years		69 (61-75)	70 (61-77)	71 (64-78)	65 (59-71)
Sex					
	Female	402 (51%)	689 (42%)	421 (46%)	477 (43%)
	Male	426 (49%)	956 (58%)	499 (54%)	645 (57%)
Stage					
	I	101 (12%)	102 (6%)	70 (8%)	
	II	317 (38%)	797 (48%)	354 (38%)	402 (36%)
	III	410 (50%)	746 (45%)	496 (54%)	720 (64%)
pN stage					
	pN0	415 (50%)	891 (54%)	425 (46%)	402 (36%)
	pN1	241 (29%)	492 (30%)	258 (28%)	508 (45%)
	pN2	167 (20%)	239 (15%)	237 (26%)	183 (16%)
	Missing	5 (1%)	23 (1%)	0 (0%)	29 (3%)
pT stage					
	pT1	26 (3%)	30 (2%)	6 (1%)	17 (2%)
	pT2	110 (13%)	137 (8%)	65 (7%)	71 (6%)
	pT3	464 (56%)	1034 (63%)	411 (45%)	582 (52%)
	pT4	223 (27%)	423 (26%)	437 (48%)	404 (36%)
	Missing	5 (1%)	21 (1%)	1 (0%)	48 (4%)
Histological grade					
	1	77 (9%)	196 (12%)	134 (15%)	45 (4%)
	2	568 (69%)	1151 (70%)	489 (53%)	846 (75%)
	3	178 (21%)	280 (17%)	297 (32%)	168 (15%)
	Missing	5 (1%)	18 (1%)	0 (0%)	63 (6%)
Location					
	Rectum	222 (27%)	457 (28%)	311 (34%)	165 (15%)
	Distal colon	262 (32%)	533 (32%)	280 (30%)	451 (40%)
	Proximal colon	307 (37%)	505 (31%)	329 (36%)	453 (40%)
	Missing	37 (4%)	150 (9%)	0 (0%)	53 (5%)
Adjuvant treatment					
	No	467 (56%)	826 (50%)	538 (58%)	0 (0%)
	Chemotherapy	173 (21%)	397 (24%)	51 (6%)	1122 (100%)
	Radiotherapy	11 (1%)	6 (0%)	14 (2%)	0 (0%)
	Chemo- and radiotherapy	3 (0%)	9 (1%)	3 (0%)	0 (0%)
	Missing	174 (21%)	407 (25%)	314 (34%)	0 (0%)
Follow-up time, years		6·4 (1·7-8·2)	4·0 (2·2-5·2)	2·4 (1·0-4·6)	4·6 (3·3-5·1)

Data are median (IQR) or number (%). IQR=interquartile range.

Table 2

Table 2: **Multivariable cancer-specific survival analyses in the validation cohort; the multivariable model included the DoMore-v1-CRC class evaluated on Aperio AT2 slide images, and established prognostic markers that were significant in the corresponding stage-specific univariable analyses in the validation cohort**

	Group	Stage II and III		Stage II		Stage III	
		HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
DoMore-v1-CRC			<0·0001		0·028		0·0001
	Good prognosis	ref.		ref.		ref.	
	Uncertain	1·56 (0·92-2·65)		1·22 (0·35-4·24)		2·14 (1·15-3·99)	
	Poor prognosis	3·04 (2·07-4·47)		2·71 (1·25-5·86)		2·95 (1·81-4·82)	
pN stage			<0·0001				<0·0001
	pN0	ref.					
	pN1	1·84 (1·13-2·98)				ref.	
	pN2	5·94 (3·71-9·52)				3·31 (2·14-5·13)	
pT stage			0·0058				0·014
	pT1	NA				NA	
	pT2	1·86 (0·90-3·86)				1·68 (0·64-4·45)	
	pT3	ref.				ref.	
	pT4	1·75 (1·22-2·51)				2·07 (1·33-3·22)	
Lymphatic invasion	Yes	1·66 (1·07-2·56)	0·023			1·98 (1·20-3·28)	0·0079
Venous vascular invasion	Yes	1·07 (0·76-1·51)	0·71			0·98 (0·64-1·52)	0·94
Sidedness	Right					1·09 (0·70-1·70)	0·69
BRAF	Mutated					1·39 (0·81-2·40)	0·24

Ref.=reference; NA=not available

Table 3

Table 3: **Associations between the DoMore-v1-CRC class evaluated on Aperio AT2 slide images and different patient characteristics in the validation cohort**

	Group	DoMore-v1-CRC good prognosis	DoMore-v1-CRC uncertain	DoMore-v1-CRC poor prognosis	Spearman's correlation	
		(N=704)	(N=136)	(N=270)	ρ (95% CI)	p
Age (continuous), years		64 (58-71)	65 (60-71)	66 (60-72)	0·07 (0·01 to 0·13)	0·024
Age (dichotomous), years					0·03 (-0·03 to 0·09)	0·38
	≤72	568 (81%)	112 (82%)	209 (77%)		
	>72	136 (19%)	24 (18%)	61 (23%)		
Sex					-0·02 (-0·08 to 0·04)	0·59
	Female	297 (42%)	53 (39%)	122 (45%)		
	Male	407 (58%)	83 (61%)	148 (55%)		
Stage					0·04 (-0·02 to 0·10)	0·20
	II	261 (37%)	48 (35%)	88 (33%)		
	III	443 (63%)	88 (65%)	182 (67%)		
Stage with substage					0·15 (0·09 to 0·21)	<0·0001
	IIA	143 (21%)	19 (14%)	28 (11%)		
	IIB	110 (16%)	27 (20%)	54 (21%)		
	IIIA	67 (10%)	2 (2%)	6 (2%)		
	IIIB	269 (40%)	51 (38%)	104 (41%)		
	IIIC	83 (12%)	34 (26%)	64 (25%)		
pN stage					0·10 (0·04 to 0·16)	0·0008
	pN0	261 (38%)	48 (36%)	88 (33%)		
	pN1	339 (50%)	53 (39%)	111 (42%)		
	pN2	83 (12%)	34 (25%)	64 (24%)		
pT stage					0·26 (0·21 to 0·32)	<0·0001
	pT1	15 (2%)	0 (0%)	2 (1%)		
	pT2	61 (9%)	3 (2%)	6 (2%)		
	pT3	402 (60%)	75 (56%)	100 (39%)		
	pT4	194 (29%)	56 (42%)	148 (58%)		
Lymphatic invasion					0·04 (-0·02 to 0·10)	0·20
	No	599 (91%)	122 (92%)	220 (87%)		
	Yes	62 (9%)	10 (8%)	33 (13%)		
Venous vascular invasion					0·05 (-0·01 to 0·11)	0·11
	No	409 (61%)	74 (56%)	145 (56%)		
	Yes	257 (39%)	58 (44%)	112 (44%)		
Histological grade					0·14 (0·08 to 0·20)	<0·0001
	1	27 (4%)	7 (6%)	8 (3%)		
	2	565 (85%)	88 (69%)	186 (74%)		
	3	76 (11%)	32 (25%)	59 (23%)		
Location					0·15 (0·09 to 0·21)	<0·0001
	Rectum	118 (18%)	21 (16%)	23 (9%)		
	Distal colon	301 (45%)	46 (35%)	100 (38%)		
	Proximal colon	246 (37%)	64 (49%)	138 (53%)		
Sidedness					0·14 (0·08 to 0·20)	<0·0001
	Left	419 (63%)	67 (51%)	123 (47%)		
	Right	246 (37%)	64 (49%)	138 (53%)		
KRAS					-0·06 (-0·12 to 0·00)	0·069
	Wild-type	410 (65%)	86 (73%)	169 (70%)		
	Mutated	224 (35%)	32 (27%)	73 (30%)		
BRAF					0·22 (0·16 to 0·28)	<0·0001
	Wild-type	588 (93%)	89 (75%)	190 (77%)		

	Mutated	47 (7%)	29 (25%)	56 (23%)		
Microsatellite instability					-0·10 (-0·16 to -0·04)	0·0018
	Yes	66 (10%)	26 (21%)	40 (16%)		
	No	595 (90%)	99 (79%)	213 (84%)		
Follow-up time, years		4·8 (3·7-5·1)	4·9 (3·1-5·1)	4·1 (2·8-5·1)	-0·10 (-0·16 to -0·04)	0·0006

Data are median (IQR) or number (%). IQR=interquartile range.

Appendix with protocol

Click here to download Supplementary Material: appendix_with_protocol.pdf

Video

Click here to download Video: THELANCET-D-19-03766.mp4

Table of Contents

Search criteria used in “Evidence before this study”

In the “Research in context” panel, the PubMed search outlined under “Evidence before this study” was the user query:
("deep learning" OR "machine learning") AND (prediction OR prognosis OR classification) AND (survival OR outcome) AND (cancer OR tumor OR tumour) AND (histology OR histopathology)
PubMed translated this user query into the following detailed search query:
("deep learning"[All Fields] OR "machine learning"[All Fields]) AND (prediction[All Fields] OR ("prognosis"[MeSH Terms] OR "prognosis"[All Fields]) OR ("classification"[Subheading] OR "classification"[All Fields] OR "classification"[MeSH Terms])) AND (("mortality"[Subheading] OR "mortality"[All Fields] OR "survival"[All Fields] OR "survival"[MeSH Terms]) OR outcome[All Fields]) AND (("neoplasms"[MeSH Terms] OR "neoplasms"[All Fields] OR "cancer"[All Fields]) OR ("tumour"[All Fields] OR "neoplasms"[MeSH Terms] OR "neoplasms"[All Fields] OR "tumor"[All Fields]) OR ("tumour"[All Fields] OR "neoplasms"[MeSH Terms] OR "neoplasms"[All Fields] OR "tumor"[All Fields])) AND (("anatomy and histology"[Subheading] OR ("anatomy"[All Fields] AND "histology"[All Fields]) OR "anatomy and histology"[All Fields] OR "histology"[All Fields] OR "histology"[MeSH Terms]) OR ("pathology"[Subheading] OR "pathology"[All Fields] OR "histopathology"[All Fields] OR "pathology"[MeSH Terms] OR "histopathology"[All Fields]))

Image classification

What follows is a description of a general framework for classifying images using multiple instance learning. The DoMore v1 network presented in the main text is a particular version of this general network architecture. Similar architectures in the context of image classification with multiple instance learning are described in the literature,[1,2] but particulars of the training process described below has not been reported elsewhere to our knowledge.

Overview
The purpose of the method is to classify an image, and involves partitioning the original image into a number of smaller patches, called tiles. The collection of all tiles in an image is denoted I. A collection of tiles from the same image is called a bag, and a collection of bags is called a batch (or mini-batch). None of the individual tiles are assigned a label, instead the bag of tiles inherits the label of the image from where it originates. We will denote a bag as a collection of tiles, $B \subseteq I$.

An integral part of this multiple instance learning method is an artificial neural network consisting of main parts: a representation network, a pooling function, and a classification network, all of which can be selected independently to fit a particular task. One update step of the training is listed below

1. A batch of bags are input to the network
2. The representation network maps each tile to a representation of the tile
3. Representations are aggregated by the pooling function
4. A classification network takes pooled representations as input and produce a prediction
5. This prediction is compared with a reference classification using some loss function
6. Derivatives of the loss function with respect to the parameters of the network is used to update the respective parameters

The representation network, the pooling function, and the final classification network, all can have trainable parameters, and the entire network is trained end-to-end.

In the rest of the description, we will ignore the batch dimension (and implicitly assume a batch size of one). Extending to a batch size larger than one works like one would expect in a regular deep learning setting with neural networks.

Representation network
The representation network is a function f_r: $\mathbb{R}^{\mathrm{m}\times\mathrm{n}\times\mathrm{c}} \to \mathbb{R}^{\mathrm{s}}$ that maps a tile x with shape $m \times n \times c$ to some feature representation of the tile $f_r(x;\ \theta_r)$ with size s. This function can for example be a regular convolutional neural network. The trainable parameters associated with the representation network are denoted θ_r.

The representation network is applied on all tiles in a bag B, producing a bag of representations $R = \{f_r(x;\ \theta_r): x\ \in\ B\}$. Note that within the same update, all tiles in a bag, and all bags in a batch uses the exact same representation network with the same values of θ_r. All representations within a batch have to be computed, and stored, before the next step.

Pooling function
The pooling function reduce the set of tile representations R to a single representation for one bag B, and is typically a function f_p: $\mathbb{R}^{\mathrm{b}\times\mathrm{s}} \to \mathbb{R}^{\mathrm{t}}$, where b is the number of tiles in a bag. Since this function potentially is dependent on the final representations of all tiles in a bag, it cannot be computed before all those representations are computed. This function can also have trainable parameters, the collection of which is denoted θ_p.

Classification network
The final part of the network is a classification network f_c: $\mathbb{R}^{\mathrm{t}} \to \mathbb{R}^{\mathrm{k}}$, where k is the number of classes. This function is parameterised with its own set of trainable parameters θ_c, the output range is typically $[0, 1]$ and such that $\sum_{i=1}^{k} f_c(x;\ \theta_c)_i = 1$ for all fitting inputs $x\ \in \mathbb{R}^{\mathrm{t}}$. With this, the output of this function can be interpreted as a prediction probability over the possible output classes, conditioned on the input. Note that all tiles in a bag contributes to one single prediction per bag, and we therefore do not get a per-tile prediction, but a per-bag prediction.

Training
The full network $f\colon \mathbb{R}^{\mathrm{b}\times\mathrm{m}\times\mathrm{n}\times\mathrm{c}} \to \mathbb{R}^{\mathrm{k}}$ produce a prediction $f(B;\ \theta_r, \theta_p, \theta_c)$ for each bag of tiles B, and this prediction is compared with a reference label assigned to the bag using a loss function L. In the following we assume that the loss function is sufficiently differentiable to be optimised using a gradient-based optimisation method.

Tile sampling
Ideally, one would like the tiles in a bag to span the entire image, but hardware constraints often necessitates subsampling. In principle, one can sample tiles from an image in many different ways, but assuming no prior knowledge, a uniform random sampling without replacement is sufficient. With a random subsampling of tiles, it is unlikely that an image will be represented by the same configuration of tiles each time. This could have a regularising effect on the training, and help generalisation.

A bag is given the label of its origin image, and if the tiles in a bag does not span the entire image, this assignment is not entirely justified. The assumption is, however, that the error made in assigning the image label to a bag of tiles is smaller than assigning the image label to a single tile. Implicit in this assumption is that the approximation error decreases with an increasing area represented by the tiles in a bag, ranging from a bag with one single tile to a bag containing all tiles in an image.

Truncated gradient contribution
It is desirable to use as many tiles as possible to represent an image in an update step of the optimisation, and for large images, the number of tiles per bag is limited by the memory of the hardware the method runs on. The representation network is the largest consumer of memory in this framework. In the forward propagation, all tiles in a bag are processed by the representation network, but only a representation of the tiles, with a considerable smaller size, is used further in the forward propagation. A gradient-based optimisation method makes use of intermediate representations of each tile "within" the network to update the parameters of the network. This means that these intermediate representations are stored until the relevant gradients are computed. By reducing the number of tiles used in the backpropagation, we would significantly reduce the memory footprint. The proposed method is to use the entire bag B in the forward propagation of the representation network, but only a subset $G \subseteq B$ of the bag in the backward propagation. Note that it is only the representation network that employ this truncation of gradient contributions. All tile representations from a bag are used by the pooling function and therefore by the final classification network, and the update of parameters associated with the pooling function and the classification network is not affected with the truncation in the representation network. It is hypothesised that increasing the size of G with B fixed will aid the optimisation. It is also hypothesised that increasing the size of B with G fixed will aid the optimisation.

Inference
In order to classify an image with a trained network, the image is tiled, and the representation network is applied on all tiles in the image. This can be done on one tile at the time, and each tile representation is stored until all tiles are processed by the representation network. Each tile representation is very small, so the number of tiles per image in inference is for all practical purposes almost limitless with respect to memory. The tile representations are aggregated by the pooling function, and the classification network produces a classification for the entire image.

Even though the bag size often is different in training and inference, a successfully trained network seems to produce reasonable results. Since the network uses all tiles in an image for inference, an image is usually better represented in inference than in training. The difficulty is often to make the network learn features that can be generalised over the entire image, which is one of the reasons why it is important with a large bag size.

Example
In the method presented in the main manuscript, the following values are used. The representation network f_r is MobileNet v2,[3] and the pooling function f_p is NoisyAND.[1] The classification network f_c is an ordinary fully connected neural network. The tile representation size, *s* is 2, and the number of inputs to the final classification network, *t* is also 2. Finally, the number of classes, *k*, is 2. In training the network, we use a batch size of 32, a bag size |*B*| of 64 and the number of tiles contributing to the gradient approximation |*G*| is 8, with an input tile size of 448x448x3. For a more detailed description, see section 5.1 in the protocol.

DoMore-v1-CRC classifier with five risk groups

In a post-hoc analysis, the continuous DoMore-v1-CRC score was categorised into five risk groups. This classifier was designed by computing the c-index of the categorised DoMore-v1-CRC score in the tuning cohort for all possible combination of four thresholds with values 0·01, 0·02, and so on up to 0·99, filtering in the c-index space by a 61 elements wide Gaussian kernel with standard deviation 0·1, and selecting the threshold combination that maximised the filtered c-index. The selected thresholds defined five risk groups by categorising the DoMore-v1-CRC score in [0, 0·39], (0·39, 0·51], (0·51, 0·61], (0·61, 0·72], and (0·72, 1].

Appendix references

1. Kraus OZ, Ba JL, Frey BJ. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics* 2016;**32**:i52–i9.
2. Carbonneau M-A, Cheplygina V, Granger E, Gagnon G. Multiple Instance Learning: A Survey of Problem Characteristics and Applications. *Pattern Recogn* 2018;**77**:329–53.
3. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2018: 4510–20.

Table S1: REporting recommendations for tumour MARKer prognostic studies (REMARK) checklist

	Item to be reported	Where reported	Comments
INTRODUCTION			
1	State the marker examined, the study objectives, and any pre-specified hypotheses.	Introduction, Methods, appendix	
MATERIALS AND METHODS			
Patients			
2	Describe the characteristics (e.g., disease stage or co-morbidities) of the study patients, including their source and inclusion and exclusion criteria.	Methods	
3	Describe treatments received and how chosen (e.g., randomized or rule-based).	Methods	
Specimen characteristics			
4	Describe type of biological material used (including control samples) and methods of preservation and storage.	Methods, appendix	
Assay methods			
5	Specify the assay method used and provide (or reference) a detailed protocol, including specific reagents or kits used, quality control procedures, reproducibility assessments, quantitation methods, and scoring and reporting protocols. Specify whether and how assays were performed blinded to the study endpoint.	Methods	Standard H&E stained sections.
Study design			
6	State the method of case selection, including whether prospective or retrospective and whether stratification or matching (e.g., by stage of disease or age) was used. Specify the time period from which cases were taken, the end of the follow-up period, and the median follow-up time.	Methods	
7	Precisely define all clinical endpoints examined.	Methods	
8	List all candidate variables initially examined or considered for inclusion in models.	Protocol	
9	Give rationale for sample size; if the study was designed to detect a specified effect size, give the target power and effect size.		Included as many samples as possible to represent variation.
Statistical analysis methods			
10	Specify all statistical methods, including details of any variable selection procedures and other model-building issues, how model assumptions were verified, and how missing data were handled.	Methods, protocol	
11	Clarify how marker values were handled in the analyses; if relevant, describe methods used for cutpoint determination.	Protocol	
RESULTS			
Data			
12	Describe the flow of patients through the study, including the number of patients included in each stage of the analysis (a diagram may be helpful) and reasons for dropout. Specifically, both overall and for each subgroup extensively examined report the numbers of patients and the number of events.	Protocol	
13	Report distributions of basic demographic characteristics (at least age and sex), standard (disease-specific) prognostic variables, and tumor marker, including numbers of missing values.	Table 1	
Analysis and presentation			
14	Show the relation of the marker to standard prognostic variables.	Tables 2 and 3	
15	Present univariable analyses showing the relation between the marker and outcome, with the estimated effect (e.g., hazard ratio and survival probability). Preferably provide similar analyses for all other variables being analyzed. For the effect of a tumor marker on a time-to-event outcome, a Kaplan-Meier plot is recommended.	Results, table 2, figure 2, appendix	
16	For key multivariable analyses, report estimated effects (e.g., hazard ratio) with confidence intervals for the marker and, at least for the final model, all other variables in the model.	Table 2	
17	Among reported results, provide estimated effects with confidence intervals from an analysis in which the marker and standard prognostic variables are included, regardless of their statistical significance.	Table 2	Included variables that were significant in univariable analysis of cancer-specific survival.
18	If done, report results of further investigations, such as checking assumptions, sensitivity analyses, and internal validation.	Ok	Internally tested in new tumour blocks prepared at a different pathology laboratory, then independently validated in a clinical trial cohort (QUASAR 2).
DISCUSSION			
19	Interpret the results in the context of the pre-specified hypotheses and other relevant studies; include a discussion of limitations of the study.	Discussion	
20	Discuss implications for future research and clinical value.	Discussion	

Table S2: Checklist for developing and reporting machine learning predictive models in biomedical research proposed by Luo et al

Item number	Topic	Checklist item
1	Nature of the study	Ok, main manuscript (Title).
2	Structured summary	Ok, main manuscript (Abstract).
3	Rationale	Ok, main manuscript (Introduction).
4	Objectives	Ok, main manuscript (Introduction), although predictive modelling commonly refers to prediction based on multiple relevant variables, whereas our study reports on the development and independent validation of a single marker which we propose to use in combination with established clinicopathological parameters.
5	Describe the setting	Ok, main manuscript ("Evidence before this study" and Introduction) and protocol.
6	Define the prediction problem	Ok, main manuscript (Introduction and Methods) and protocol. Development and independent validation of a prognostic marker in retrospective datasets. Not all fields are applicable to our study.
7	Prepare data for model building	Ok, protocol.
8	Build the predictive model	Ok, main manuscript (Methods) and protocol. Not all fields are applicable to our study.
9	Report the final model and performance	Ok, main manuscript (tables 2 and 3) and appendix.
10	Clinical implications	Ok, main manuscript ("Added value of this study" and Discussion).
11	Limitations of the model	Ok, main manuscript (Discussion).
12	Unexpected results during the experiments	Not applicable.

Table S3: Univariable cancer-specific survival analyses in the validation cohort of the DoMore-v1-CRC class and score, its constituents, and established prognostic markers

Variable	Group	Stage II and III		Stage II		Stage III	
		HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
DoMore-v1-CRC class on Aperio AT2			<0·0001		0·028		<0·0001
	Good prognosis	ref.		ref.		ref.	
	Uncertain	1·89 (1·14-3·15)		1·22 (0·35-4·24)		2·07 (1·18-3·63)	
	Poor prognosis	3·84 (2·72-5·43)		2·71 (1·25-5·86)		4·09 (2·77-6·03)	
10x ensemble class on Aperio AT2	Poor prognosis	3·26 (2·37-4·49)	<0·0001	2·42 (1·16-5·08)	0·015	3·40 (2·38-4·85)	<0·0001
40x ensemble class on Aperio AT2	Poor prognosis	3·19 (2·30-4·42)	<0·0001	2·32 (1·12-4·80)	0·020	3·38 (2·34-4·88)	<0·0001
DoMore-v1-CRC score on Aperio AT2	25% increment	2·48 (1·98-3·11)	<0·0001	2·02 (1·19-3·44)	0·0095	2·54 (1·97-3·27)	<0·0001
10x ensemble score on Aperio AT2	25% increment	2·16 (1·80-2·61)	<0·0001	1·84 (1·18-2·85)	0·0070	2·20 (1·79-2·70)	<0·0001
40x ensemble score on Aperio AT2	25% increment	2·69 (2·05-3·53)	<0·0001	2·09 (1·12-3·92)	0·021	2·80 (2·07-3·80)	<0·0001
DoMore-v1-CRC class on NanoZoomer XR			<0·0001		0·0021		<0·0001
	Good prognosis	ref.		ref.		ref.	
	Uncertain	2·42 (1·45-4·03)		2·78 (0·84-9·25)		2·22 (1·26-3·91)	
	Poor prognosis	3·39 (2·36-4·87)		4·00 (1·74-9·20)		3·20 (2·14-4·80)	
10x ensemble class on NanoZoomer XR	Poor prognosis	3·34 (2·38-4·68)	<0·0001	3·89 (1·78-8·49)	0·0002	3·18 (2·19-4·63)	<0·0001
40x ensemble class on NanoZoomer XR	Poor prognosis	2·46 (1·78-3·40)	<0·0001	2·87 (1·37-6·04)	0·0035	2·31 (1·61-3·32)	<0·0001
DoMore-v1-CRC score on NanoZoomer XR	25% increment	2·51 (1·98-3·19)	<0·0001	2·64 (1·54-4·50)	0·0004	2·44 (1·87-3·18)	<0·0001
10x ensemble score on NanoZoomer XR	25% increment	2·18 (1·80-2·65)	<0·0001	2·26 (1·46-3·50)	0·0003	2·12 (1·71-2·64)	<0·0001
40x ensemble score on NanoZoomer XR	25% increment	2·73 (2·04-3·64)	<0·0001	2·93 (1·52-5·67)	0·0014	2·65 (1·93-3·64)	<0·0001
pN stage			<0·0001				<0·0001
	pN0	ref.					
	pN1	1·49 (0·95-2·32)				ref.	
	pN2	6·18 (4·00-9·54)				4·15 (2·89-5·96)	
pT stage			<0·0001		0·96		<0·0001
	pT1	n/a				n/a	
	pT2	1·34 (0·66-2·71)				1·18 (0·58-2·42)	
	pT3	ref.		ref.		ref.	
	pT4	2·19 (1·56-3·07)		1·02 (0·49-2·14)		3·39 (2·31-4·98)	
Lymphatic invasion	Yes	1·87 (1·22-2·89)	0·0037	0·34 (0·05-2·50)	0·27	2·33 (1·49-3·65)	0·0001
Venous vascular invasion	Yes	1·48 (1·07-2·05)	0·018	1·11 (0·52-2·37)	0·78	1·74 (1·21-2·50)	0·0023
Age at randomisation	10-year increment	1·12 (0·95-1·33)	0·19	0·94 (0·65-1·37)	0·76	1·13 (0·93-1·36)	0·21
Sex	Male	1·12 (0·81-1·55)	0·49	1·38 (0·65-2·96)	0·40	1·09 (0·76-1·57)	0·63
Histological grade			0·24		0·91		0·057
	1	ref.		ref.		ref.	
	2	1·23 (0·50-3·02)		1·30 (0·18-9·68)		1·19 (0·44-3·24)	
	3	1·72 (0·66-4·45)		1·07 (0·12-9·55)		2·02 (0·70-5·82)	
Sidedness	Right	1·24 (0·89-1·72)	0·20	0·78 (0·37-1·67)	0·52	1·60 (1·11-2·30)	0·010
KRAS	Mutated	1·09 (0·77-1·54)	0·64	1·17 (0·55-2·51)	0·68	1·03 (0·69-1·53)	0·89
BRAF	Mutated	1·54 (0·99-2·39)	0·054	1·24 (0·47-3·23)	0·67	1·71 (1·04-2·81)	0·033
Microsatellite instability	No	1·53 (0·84-2·76)	0·16	1·89 (0·57-6·23)	0·29	1·18 (0·60-2·33)	0·64

Table S4: Multivariable cancer-specific survival analyses in the validation cohort; the multivariable model included the 10x ensemble class of the DoMore v1 network evaluated on Aperio AT2 slide images, and established prognostic markers that were significant in the corresponding stage-specific univariable analyses in the validation cohort

Variable	Group	Stage II and III		Stage II		Stage III	
		HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
10x ensemble class	Poor prognosis	2·62 (1·85-3·71)	<0·0001	2·42 (1·16-5·08)	0·015	2·40 (1·56-3·69)	0·0001
pN stage			<0·0001				<0·0001
	pN0	ref.					
	pN1	1·86 (1·15-3·02)				ref.	
	pN2	6·19 (3·88-9·89)				3·47 (2·25-5·35)	
pT stage			0·011				0·012
	pT1	n/a				n/a	
	pT2	1·76 (0·86-3·64)				1·51 (0·58-3·95)	
	pT3	ref.				ref.	
	pT4	1·80 (1·26-2·59)				2·12 (1·36-3·29)	
Lymphatic invasion	Yes	1·70 (1·10-2·63)	0·017			2·00 (1·21-3·31)	0·0069
Venous vascular invasion	Yes	1·08 (0·77-1·53)	0·66			1·02 (0·66-1·57)	0·92
Sidedness	Right					1·09 (0·70-1·69)	0·70
BRAF	Mutated					1·43 (0·83-2·46)	0·19

Table S5: Multivariable cancer-specific survival analyses in the validation cohort; the multivariable model included the 40x ensemble class of the DoMore v1 network evaluated on Aperio AT2 slide images, and established prognostic markers that were significant in the corresponding stage-specific univariable analyses in the validation cohort

Variable	Group	Stage II and III		Stage II		Stage III	
		HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
40x ensemble class	Poor prognosis	2·51 (1·75-3·59)	<0·0001	2·32 (1·12-4·80)	0·020	2·59 (1·64-4·08)	<0·0001
pN stage			<0·0001				<0·0001
	pN0	ref.					
	pN1	1·88 (1·16-3·05)				ref.	
	pN2	5·82 (3·63-9·32)				3·22 (2·08-4·98)	
pT stage			0·0046				0·0068
	pT1	n/a				n/a	
	pT2	1·85 (0·89-3·85)				1·66 (0·63-4·37)	
	pT3	ref.				ref.	
	pT4	1·88 (1·32-2·69)				2·17 (1·40-3·35)	
Lymphatic invasion	Yes	1·63 (1·06-2·53)	0·027			1·96 (1·18-3·24)	0·0089
Venous vascular invasion	Yes	1·08 (0·76-1·52)	0·67			0·98 (0·64-1·51)	0·93
Sidedness	Right					1·12 (0·72-1·74)	0·61
BRAF	Mutated					1·39 (0·81-2·41)	0·23

Table S6: Areas under the curve (AUC) for patients with distinct outcome and Harrell's concordance index for all patients in the validation cohort, both with 95% confidence intervals (CIs), between cancer-specific survival and a DoMore v1 individual model score

Variable	AUC (95% CI)	c-index (95% CI)
10x model 1 score on Aperio AT2	0·713 (0·624-0·790)	0·681 (0·632-0·724)
10x model 2 score on Aperio AT2	0·702 (0·610-0·779)	0·671 (0·620-0·716)
10x model 3 score on Aperio AT2	0·705 (0·614-0·782)	0·662 (0·612-0·708)
10x model 4 score on Aperio AT2	0·680 (0·591-0·757)	0·656 (0·606-0·702)
10x model 5 score on Aperio AT2	0·740 (0·651-0·815)	0·676 (0·626-0·720)
40x model 1 score on Aperio AT2	0·681 (0·589-0·760)	0·660 (0·612-0·707)
40x model 2 score on Aperio AT2	0·682 (0·592-0·762)	0·634 (0·585-0·680)
40x model 3 score on Aperio AT2	0·686 (0·594-0·763)	0·648 (0·599-0·695)
40x model 4 score on Aperio AT2	0·697 (0·605-0·776)	0·660 (0·611-0·707)
40x model 5 score on Aperio AT2	0·711 (0·618-0·785)	0·671 (0·623-0·718)
10x model 1 score on NanoZoomer XR	0·702 (0·612-0·779)	0·670 (0·621-0·714)
10x model 2 score on NanoZoomer XR	0·707 (0·619-0·784)	0·670 (0·619-0·715)
10x model 3 score on NanoZoomer XR	0·720 (0·635-0·795)	0·664 (0·614-0·707)
10x model 4 score on NanoZoomer XR	0·712 (0·626-0·787)	0·667 (0·616-0·710)
10x model 5 score on NanoZoomer XR	0·723 (0·633-0·800)	0·672 (0·621-0·718)
40x model 1 score on NanoZoomer XR	0·700 (0·612-0·776)	0·650 (0·602-0·696)
40x model 2 score on NanoZoomer XR	0·665 (0·575-0·746)	0·631 (0·584-0·677)
40x model 3 score on NanoZoomer XR	0·696 (0·607-0·772)	0·647 (0·600-0·693)
40x model 4 score on NanoZoomer XR	0·682 (0·590-0·759)	0·654 (0·604-0·699)
40x model 5 score on NanoZoomer XR	0·698 (0·611-0·774)	0·667 (0·618-0·712)

Table S7: Multivariable cancer-specific survival analyses in the validation cohort; the multivariable model included the average of the two DoMore v1 ensemble scores evaluated on Aperio AT2 slide images, and established prognostic markers that were significant in the corresponding stage-specific univariable analyses in the validation cohort

Variable	Group	Stage II and III		Stage II		Stage III	
		HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
DoMore-v1-CRC score	25% increment	2·11 (1·63-2·73)	<0·0001	2·02 (1·19-3·44)	0·0095	1·97 (1·42-2·73)	<0·0001
pN stage			<0·0001				<0·0001
	pN0	ref.					
	pN1	1·78 (1·10-2·89)				ref.	
	pN2	5·87 (3·68-9·38)				3·46 (2·25-5·32)	
pT stage			0·012				0·017
	pT1	n/a				n/a	
	pT2	2·04 (0·98-4·23)				1·74 (0·66-4·60)	
	pT3	ref.				ref.	
	pT4	1·74 (1·21-2·50)				2·03 (1·30-3·17)	
Lymphatic invasion	Yes	1·70 (1·10-2·62)	0·018			2·06 (1·25-3·40)	0·0046
Venous vascular invasion	Yes	1·03 (0·73-1·46)	0·86			0·99 (0·64-1·52)	0·96
Sidedness	Right					1·09 (0·70-1·69)	0·71
BRAF	Mutated					1·34 (0·77-2·31)	0·30

Table S8: Multivariable cancer-specific survival analyses in the validation cohort; the multivariable model included the 10x ensemble score of the DoMore v1 network evaluated on Aperio AT2 slide images, and established prognostic markers that were significant in the corresponding stage-specific univariable analyses in the validation cohort

Variable	Group	Stage II and III		Stage II		Stage III	
		HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
10x ensemble score	25% increment	1·90 (1·54-2·35)	<0·0001	1·84 (1·18-2·85)	0·0070	1·79 (1·37-2·34)	<0·0001
pN stage			<0·0001				<0·0001
	pN0	ref.					
	pN1	1·76 (1·08-2·85)				ref.	
	pN2	5·94 (3·72-9·48)				3·53 (2·29-5·44)	
pT stage			0·017				0·020
	pT1	n/a				n/a	
	pT2	2·03 (0·98-4·21)				1·75 (0·66-4·62)	
	pT3	ref.				ref.	
	pT4	1·71 (1·19-2·45)				2·01 (1·29-3·14)	
Lymphatic invasion	Yes	1·70 (1·10-2·63)	0·017			2·07 (1·26-3·41)	0·0043
Venous vascular invasion	Yes	1·03 (0·73-1·46)	0·85			0·99 (0·65-1·53)	0·98
Sidedness	Right					1·09 (0·70-1·70)	0·69
BRAF	Mutated					1·34 (0·77-2·30)	0·30

Table S9: Multivariable cancer-specific survival analyses in the validation cohort; the multivariable model included the 40x ensemble score of the DoMore v1 network evaluated on Aperio AT2 slide images, and established prognostic markers that were significant in the corresponding stage-specific univariable analyses in the validation cohort

Variable	Group	Stage II and III		Stage II		Stage III	
		HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
40x ensemble score	25% increment	2·20 (1·62-2·99)	<0·0001	2·09 (1·12-3·92)	0·021	2·08 (1·41-3·07)	0·0002
pN stage			<0·0001				<0·0001
	pN0	ref.					
	pN1	1·84 (1·14-2·99)				ref.	
	pN2	5·89 (3·69-9·42)				3·35 (2·18-5·16)	
pT stage			0·0019				0·0094
	pT1	n/a				n/a	
	pT2	1·95 (0·94-4·05)				1·68 (0·64-4·45)	
	pT3	ref.				ref.	
	pT4	1·84 (1·29-2·63)				2·13 (1·37-3·30)	
Lymphatic invasion	Yes	1·69 (1·10-2·62)	0·018			2·04 (1·23-3·36)	0·0053
Venous vascular invasion	Yes	1·04 (0·74-1·47)	0·81			0·99 (0·64-1·52)	0·96
Sidedness	Right					1·09 (0·70-1·69)	0·71
BRAF	Mutated					1·36 (0·79-2·35)	0·27

Table S10: Harrell's concordance index (95% CI) between cancer-specific survival and the DoMore-v1-CRC or Inception v3 score, or one of their constituents

Variable	Test cohort	Validation cohort
DoMore-v1-CRC score on Aperio AT2	0·695 (0·659-0·726)	0·674 (0·624-0·719)
DoMore v1 10x ensemble score on Aperio AT2	0·691 (0·654-0·722)	0·677 (0·627-0·722)
DoMore v1 40x ensemble score on Aperio AT2	0·690 (0·656-0·721)	0·664 (0·615-0·711)
Inception v3 score on Aperio AT2	0·679 (0·642-0·712)	0·654 (0·605-0·700)
Inception v3 10x ensemble score on Aperio AT2	0·663 (0·626-0·698)	0·663 (0·615-0·709)
Inception v3 40x ensemble score on Aperio AT2	0·674 (0·641-0·707)	0·628 (0·578-0·676)
DoMore-v1-CRC score on NanoZoomer XR	0·692 (0·656-0·723)	0·674 (0·624-0·718)
DoMore v1 10x ensemble score on NanoZoomer XR	0·689 (0·652-0·720)	0·678 (0·628-0·722)
DoMore v1 40x ensemble score on NanoZoomer XR	0·683 (0·649-0·715)	0·659 (0·610-0·704)
Inception v3 score on NanoZoomer XR	0·677 (0·641-0·711)	0·649 (0·598-0·695)
Inception v3 10x ensemble score on NanoZoomer XR	0·659 (0·621-0·694)	0·651 (0·602-0·696)
Inception v3 40x ensemble score on NanoZoomer XR	0·680 (0·646-0·713)	0·634 (0·586-0·682)

Table S11: Cross-tabulation of DoMore-v1-CRC class evaluated on Aperio AT2 slide images and histological grade in the test cohort

DoMore-v1-CRC class	Well differentiated	Moderately differentiated	Poorly differentiated
Good prognosis	101 (76%)	175 (36%)	33 (11%)
Uncertain	18 (14%)	106 (22%)	36 (12%)
Poor prognosis	14 (11%)	199 (41%)	223 (76%)

Table S12: Multivariable cancer-specific survival analyses in the validation cohort; the multivariable model included the DoMore-v1-CRC class evaluated on NanoZoomer XR slide images, and established prognostic markers that were significant in the corresponding stage-specific univariable analyses in the validation cohort

Variable	Group	Stage II and III		Stage II		Stage III	
		HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
DoMore-v1-CRC			0·0001		0·0021		0·019
	Good prognosis	ref.		ref.		ref.	
	Uncertain	1·80 (1·05-3·10)		2·78 (0·84-9·25)		1·63 (0·84-3·15)	
	Poor prognosis	2·46 (1·65-3·67)		4·00 (1·74-9·20)		2·07 (1·25-3·44)	
pN stage			<0·0001				<0·0001
	pN0	ref.					
	pN1	1·89 (1·17-3·04)				ref.	
	pN2	5·51 (3·45-8·79)				3·00 (1·94-4·63)	
pT stage			0·0004				0·0017
	pT1	n/a				n/a	
	pT2	1·92 (0·92-3·99)				1·53 (0·58-4·01)	
	pT3	ref.				ref.	
	pT4	2·00 (1·41-2·85)				2·36 (1·53-3·64)	
Lymphatic invasion	Yes	1·68 (1·09-2·60)	0·019			2·03 (1·23-3·34)	0·0056
Venous vascular invasion	Yes	1·11 (0·78-1·56)	0·57			1·05 (0·68-1·62)	0·83
Sidedness	Right					1·15 (0·74-1·78)	0·54
BRAF	Mutated					1·34 (0·78-2·32)	0·29

Table S13: Multivariable cancer-specific survival analyses in the validation cohort; the multivariable model included the 10x ensemble class of the DoMore v1 network evaluated on NanoZoomer XR slide images, and established prognostic markers that were significant in the corresponding stage-specific univariable analyses in the validation cohort

Variable	Group	Stage II and III		Stage II		Stage III	
		HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
10x ensemble class	Poor prognosis	2·56 (1·77-3·70)	<0·0001	3·89 (1·78-8·49)	0·0002	2·17 (1·37-3·44)	0·0009
pN stage			<0·0001				<0·0001
	pN0	ref.					
	pN1	1·88 (1·17-3·03)				ref·	
	pN2	5·70 (3·58-9·08)				3·11 (2·02-4·80)	
pT stage			0·0020				0·0025
	pT1	n/a				n/a	
	pT2	1·86 (0·90-3·86)				1·51 (0·58-3·95)	
	pT3	ref.				ref.	
	pT4	1·96 (1·38-2·79)				2·30 (1·49-3·55)	
Lymphatic invasion	Yes	1·66 (1·07-2·56)	0·023			2·02 (1·22-3·33)	0·0060
Venous vascular invasion	Yes	1·05 (0·74-1·49)	0·77			0·99 (0·64-1·53)	0·97
Sidedness	Right					1·14 (0·74-1·77)	0·54
BRAF	Mutated					1·38 (0·80-2·38)	0·24

Table S14: Multivariable cancer-specific survival analyses in the validation cohort; the multivariable model included the 40x ensemble class of the DoMore v1 network evaluated on NanoZoomer XR slide images, and established prognostic markers that were significant in the corresponding stage-specific univariable analyses in the validation cohort

Variable	Group	Stage II and III		Stage II		Stage III	
		HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
40x ensemble class	Poor prognosis	1·75 (1·22-2·49)	0·0021	2·87 (1·37-6·04)	0·0035	1·50 (0·96-2·36)	0·075
pN stage			<0·0001				<0·0001
	pN0	ref.					
	pN1	1·95 (1·21-3·14)				ref.	
	pN2	5·77 (3·62-9·19)				3·05 (1·98-4·70)	
pT stage			0·0004				0·0005
	pT1	n/a				n/a	
	pT2	1·74 (0·84-3·61)				1·43 (0·55-3·73)	
	pT3	ref.				ref.	
	pT4	2·14 (1·50-3·04)				2·51 (1·63-3·86)	
Lymphatic invasion	Yes	1·70 (1·10-2·63)	0·016			2·03 (1·23-3·34)	0·0056
Venous vascular invasion	Yes	1·17 (0·83-1·64)	0·37			1·11 (0·72-1·69)	0·64
Sidedness	Right					1·17 (0·76-1·80)	0·49
BRAF	Mutated					1·39 (0·80-2·40)	0·24

Table S15: Multivariable cancer-specific survival analyses in the validation cohort; the multivariable model included the average of the two DoMore v1 ensemble scores evaluated on NanoZoomer XR slide images, and established prognostic markers that were significant in the corresponding stage-specific univariable analyses in the validation cohort

Variable	Group	Stage II and III		Stage II		Stage III	
		HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
DoMore-v1-CRC score	25% increment	2·05 (1·57-2·68)	<0·0001	2·64 (1·54-4·50)	0·0004	1·81 (1·29-2·53)	0·0006
pN stage			<0·0001				<0·0001
	pN0	ref.					
	pN1	1·81 (1·12-2·91)				ref.	
	pN2	5·58 (3·51-8·87)				3·22 (2·10-4·93)	
pT stage			0·0043				0·0018
	pT1	n/a				n/a	
	pT2	2·07 (0·99-4·30)				1·66 (0·63-4·39)	
	pT3	ref.				ref.	
	pT4	1·85 (1·30-2·65)				2·20 (1·42-3·41)	
Lymphatic invasion	Yes	1·66 (1·08-2·57)	0·022			2·04 (1·24-3·35)	0·0052
Venous vascular invasion	Yes	1·06 (0·75-1·50)	0·73			1·02 (0·66-1·56)	0·95
Sidedness	Right					1·11 (0·72-1·73)	0·63
BRAF	Mutated					1·37 (0·80-2·37)	0·26

Table S16: Multivariable cancer-specific survival analyses in the validation cohort; the multivariable model included the 10x ensemble score of the DoMore v1 network evaluated on NanoZoomer XR slide images, and established prognostic markers that were significant in the corresponding stage-specific univariable analyses in the validation cohort

Variable	Group	Stage II and III		Stage II		Stage III	
		HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
10x ensemble score	25% increment	1·85 (1·49-2·30)	<0·0001	2·26 (1·46-3·50)	0·0003	1·66 (1·26-2·18)	0·0003
pN stage			<0·0001				<0·0001
	pN0	ref.					
	pN1	1·77 (1·10-2·85)				ref.	
	pN2	5·58 (3·51-8·86)				3·29 (2·14-5·04)	
pT stage			0·0021				0·0070
	pT1	n/a				n/a	
	pT2	2·09 (1·01-4·36)				1·69 (0·64-4·46)	
	pT3	ref.				ref.	
	pT4	1·81 (1·26-2·59)				2·17 (1·40-3·36)	
Lymphatic invasion	Yes	1·66 (1·07-2·56)	0·023			2·02 (1·22-3·32)	0·0058
Venous vascular invasion	Yes	1·06 (0·75-1·50)	0·73			1·01 (0·66-1·56)	0·95
Sidedness	Right					1·11 (0·72-1·72)	0·64
BRAF	Mutated					1·39 (0·80-2·39)	0·24

Table S17: Multivariable cancer-specific survival analyses in the validation cohort; the multivariable model included the 40x ensemble score of the DoMore v1 network evaluated on NanoZoomer XR slide images, and established prognostic markers that were significant in the corresponding stage-specific univariable analyses in the validation cohort

Variable	Group	Stage II and III		Stage II		Stage III	
		HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
40x ensemble score	25% increment	2·11 (1·53-2·90)	<0·0001	2·93 (1·52-5·67)	0·0014	1·88 (1·25-2·82)	0·0022
pN stage			<0·0001				<0·0001
	pN0	ref.					
	pN1	1·90 (1·18-3·05)				ref.	
	pN2	5·71 (3·59-9·06)				3·13 (2·04-4·80)	
pT stage			0·0018				0·0009
	pT1	n/a				n/a	
	pT2	1·92 (0·92-3·98)				1·58 (0·60-4·16)	
	pT3	ref.				ref.	
	pT4	1·97 (1·38-2·80)				2·30 (1·49-3·55)	
Lymphatic invasion	Yes	1·69 (1·09-2·60)	0·019			2·06 (1·25-3·40)	0·0045
Venous vascular invasion	Yes	1·09 (0·77-1·53)	0·64			1·03 (0·67-1·58)	0·89
Sidedness	Right					1·13 (0·73-1·76)	0·57
BRAF	Mutated					1·37 (0·79-2·37)	0·26

Table S18: Associations between the DoMore-v1-CRC class evaluated on NanoZoomer XR slide images and different patient characteristics in the validation cohort

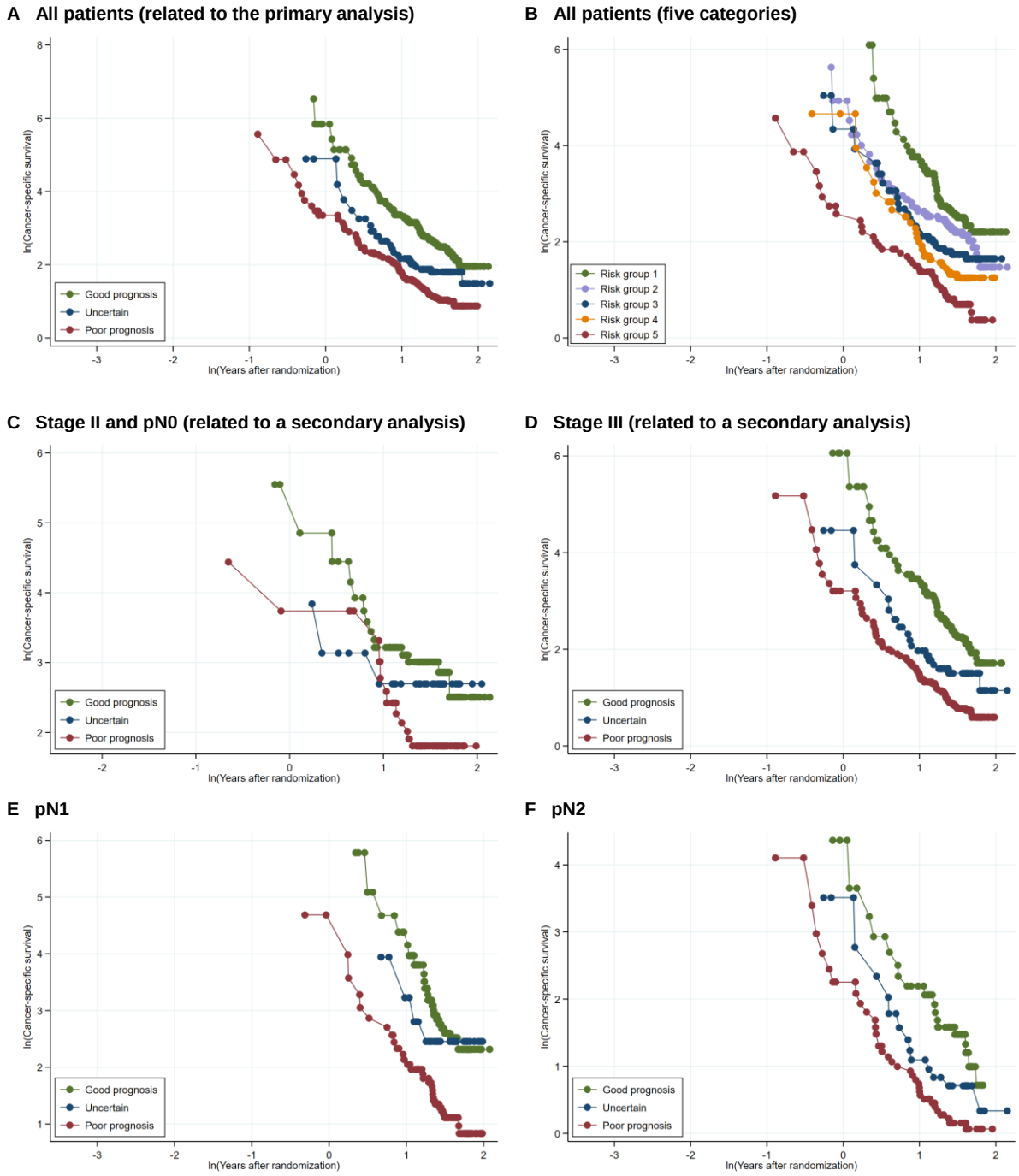
Characteristic	DoMore-v1-CRC good prognosis (N=596)	DoMore-v1-CRC uncertain (N=130)	DoMore-v1-CRC poor prognosis (N=393)	Spearman's correlation ρ (95% CI)	p
Median age at randomisation (IQR), years	64 (57-71)	67 (60-73)	66 (60-71)	0·09 (0·03 to 0·14)	0·0042
Age at randomisation, years				0·05 (-0·01 to 0·11)	0·092
≤72	494 (83%)	93 (72%)	312 (79%)		
>72	102 (17%)	37 (28%)	81 (21%)		
Sex				-0·02 (-0·08 to 0·03)	0·42
Female	245 (41%)	61 (47%)	170 (43%)		
Male	351 (59%)	69 (53%)	223 (57%)		
Stage				0·03 (-0·03 to 0·09)	0·34
II	222 (37%)	42 (32%)	136 (35%)		
III	374 (63%)	88 (68%)	257 (65%)		
Stage with substage				0·15 (0·09 to 0·21)	<0·0001
IIA	123 (22%)	21 (16%)	48 (13%)		
IIB	91 (16%)	20 (16%)	81 (21%)		
IIIA	58 (10%)	5 (4%)	12 (3%)		
IIIB	233 (41%)	51 (40%)	144 (38%)		
IIIC	59 (10%)	31 (24%)	93 (25%)		
pN stage				0·11 (0·05 to 0·16)	0·0004
pN0	222 (39%)	42 (33%)	136 (35%)		
pN1	294 (51%)	56 (43%)	157 (41%)		
pN2	59 (10%)	31 (24%)	93 (24%)		
pT stage				0·23 (0·18 to 0·29)	<0·0001
pT1	13 (2%)	2 (2%)	2 (1%)		
pT2	55 (10%)	4 (3%)	11 (3%)		
pT3	338 (60%)	69 (54%)	175 (46%)		
pT4	160 (28%)	52 (41%)	190 (50%)		
Lymphatic invasion				0·01 (-0·05 to 0·07)	0·64
No	503 (90%)	114 (93%)	331 (89%)		
Yes	56 (10%)	9 (7%)	42 (11%)		
Venous vascular invasion				0·08 (0·02 to 0·14)	0·0074
No	354 (63%)	73 (59%)	205 (54%)		
Yes	208 (37%)	51 (41%)	173 (46%)		
Histological grade				0·14 (0·08 to 0·20)	<0·0001
1	24 (4%)	8 (7%)	12 (3%)		
2	477 (85%)	94 (77%)	274 (73%)		
3	61 (11%)	20 (16%)	87 (23%)		
Location				0·11 (0·05 to 0·17)	0·0002
Rectum	97 (17%)	19 (15%)	48 (13%)		
Distal colon	254 (45%)	54 (43%)	142 (37%)		
Proximal colon	210 (37%)	52 (42%)	190 (50%)		
Sidedness				0·12 (0·06 to 0·17)	0·0002
Left	351 (63%)	73 (58%)	190 (50%)		
Right	210 (37%)	52 (42%)	190 (50%)		
KRAS				-0·08 (-0·14 to -0·02)	0·0084
Wild-type	342 (64%)	77 (66%)	252 (72%)		
Mutated	195 (36%)	40 (34%)	96 (28%)		
BRAF				0·21 (0·15 to 0·27)	<0·0001
Wild-type	502 (93%)	96 (83%)	277 (78%)		
Mutated	36 (7%)	19 (17%)	78 (22%)		
Microsatellite instability				-0·10 (-0·16 to -0·04)	0·0016
Yes	53 (9%)	21 (17%)	58 (16%)		
No	511 (91%)	101 (83%)	304 (84%)		
Median follow-up time (IQR), years	4·8 (3·8-5·1)	4·8 (3·1-5·2)	4·1 (3·1-5·1)	-0·12 (-0·18 to -0·06)	0·0001

Data are median (IQR) or number (%). IQR=interquartile range.

Table S19: Cross-tabulation of DoMore-v1-CRC class evaluated on NanoZoomer XR slide images and histological grade in the test cohort

DoMore-v1-CRC class	Well differentiated	Moderately differentiated	Poorly differentiated
Good prognosis	94 (70%)	180 (37%)	33 (11%)
Uncertain	20 (15%)	77 (16%)	38 (13%)
Poor prognosis	20 (15%)	229 (47%)	226 (76%)

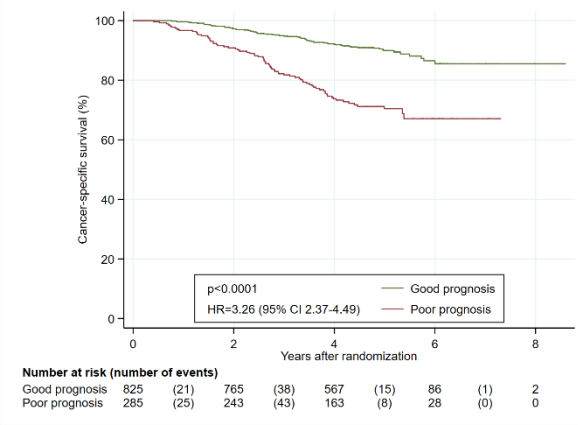
Figure S1: log-log plots of cancer-specific survival by DoMore-v1-CRC class evaluated on Aperio AT2 slide images in the validation cohort (comparable to figure 2)



The pre-defined DoMore-v1-CRC classifier was evaluated in Panels A, C, D, E, and F. The DoMore-v1-CRC classifier variant with five categories was evaluated in Panel B.

Figure S2: Kaplan-Meier analysis of cancer-specific survival by the constituents of the DoMore-v1-CRC class evaluated on Aperio AT2 slide images in the validation cohort

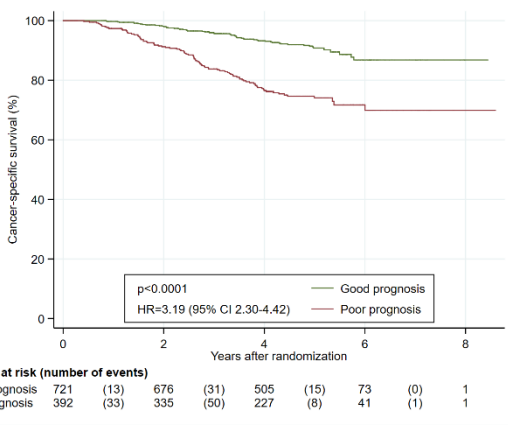
A 10x ensemble classifier (secondary analysis)



Number at risk (number of events)

Good prognosis	825	(21)	765	(38)	567	(15)	86	(1)	2
Poor prognosis	285	(25)	243	(43)	163	(8)	28	(0)	0

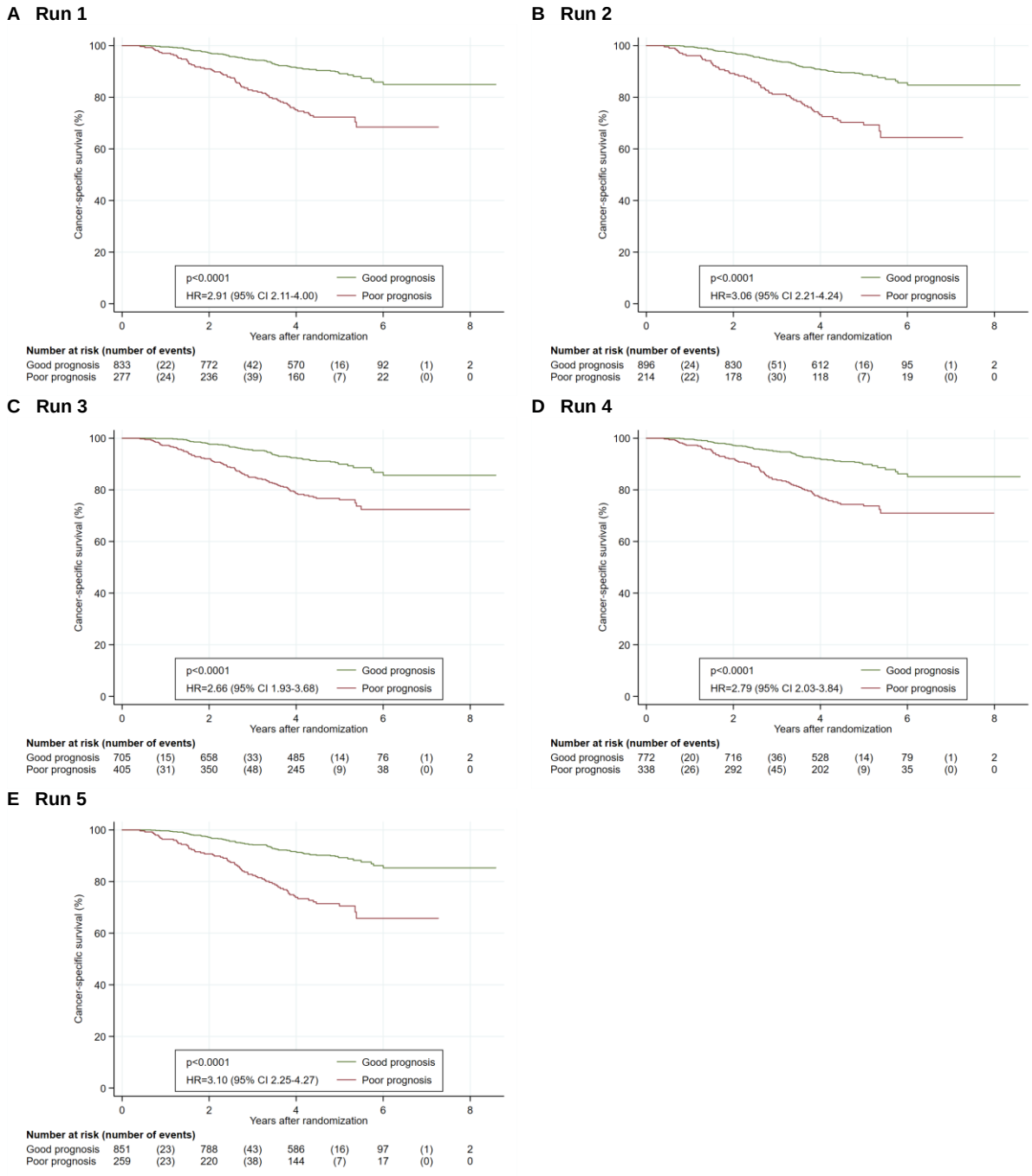
B 40x ensemble classifier (secondary analysis)



Number at risk (number of events)

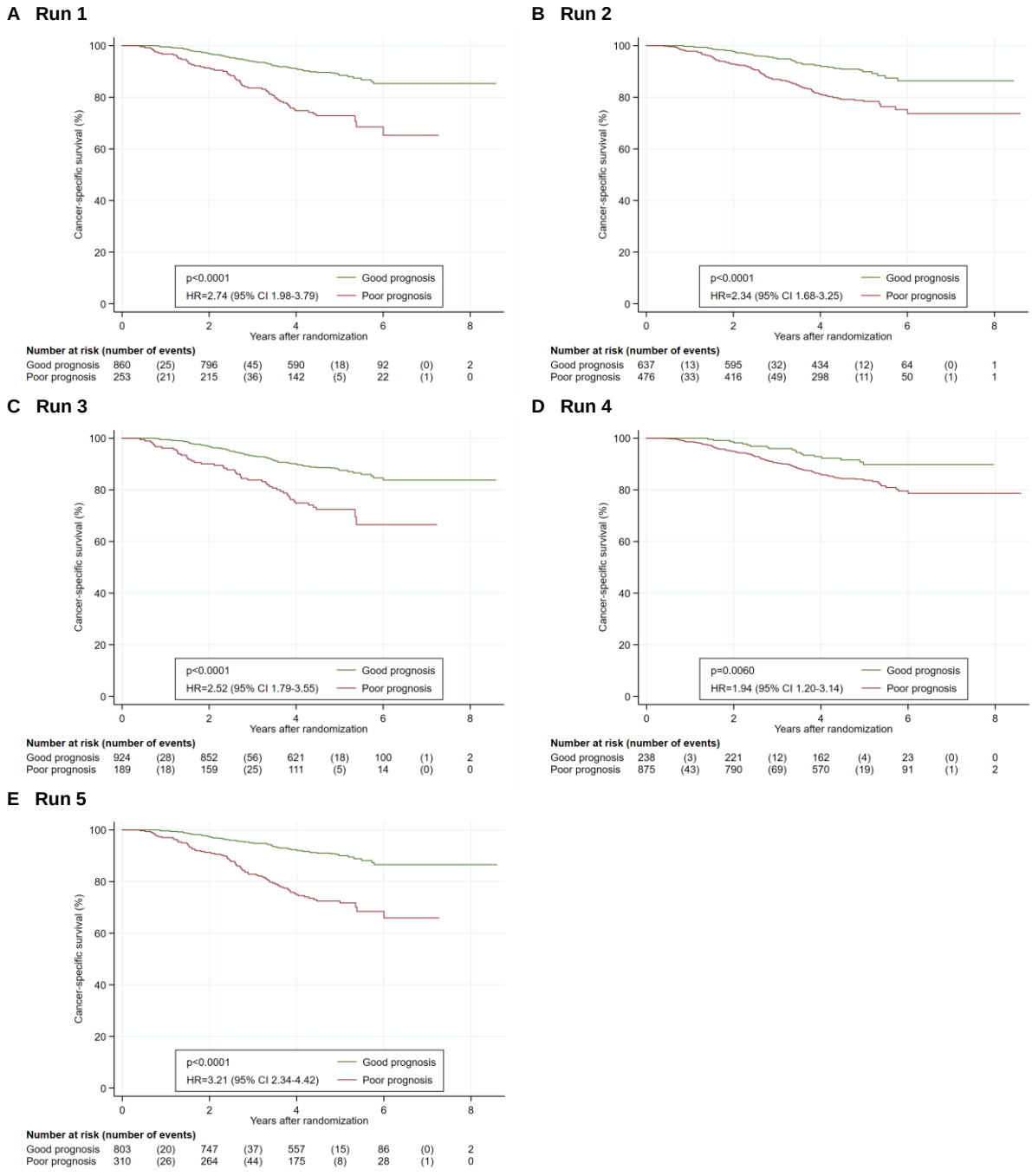
Good prognosis	721	(13)	676	(31)	505	(15)	73	(0)	1
Poor prognosis	392	(33)	335	(50)	227	(8)	41	(1)	1

Figure S3: Kaplan-Meier analysis of cancer-specific survival by the class of a DoMore v1 10x individual model evaluated on Aperio AT2 slide images in the validation cohort



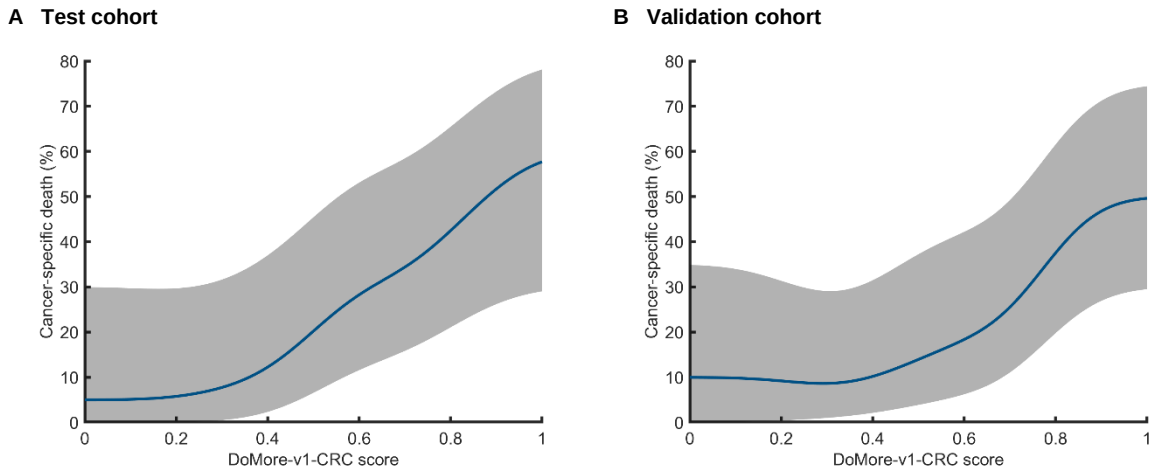
For each DoMore v1 10x individual model, the score was dichotomised using the same threshold as used for the DoMore v1 10x ensemble model.

Figure S4: Kaplan-Meier analysis of cancer-specific survival by the class of a DoMore v1 40x individual model evaluated on Aperio AT2 slide images in the validation cohort



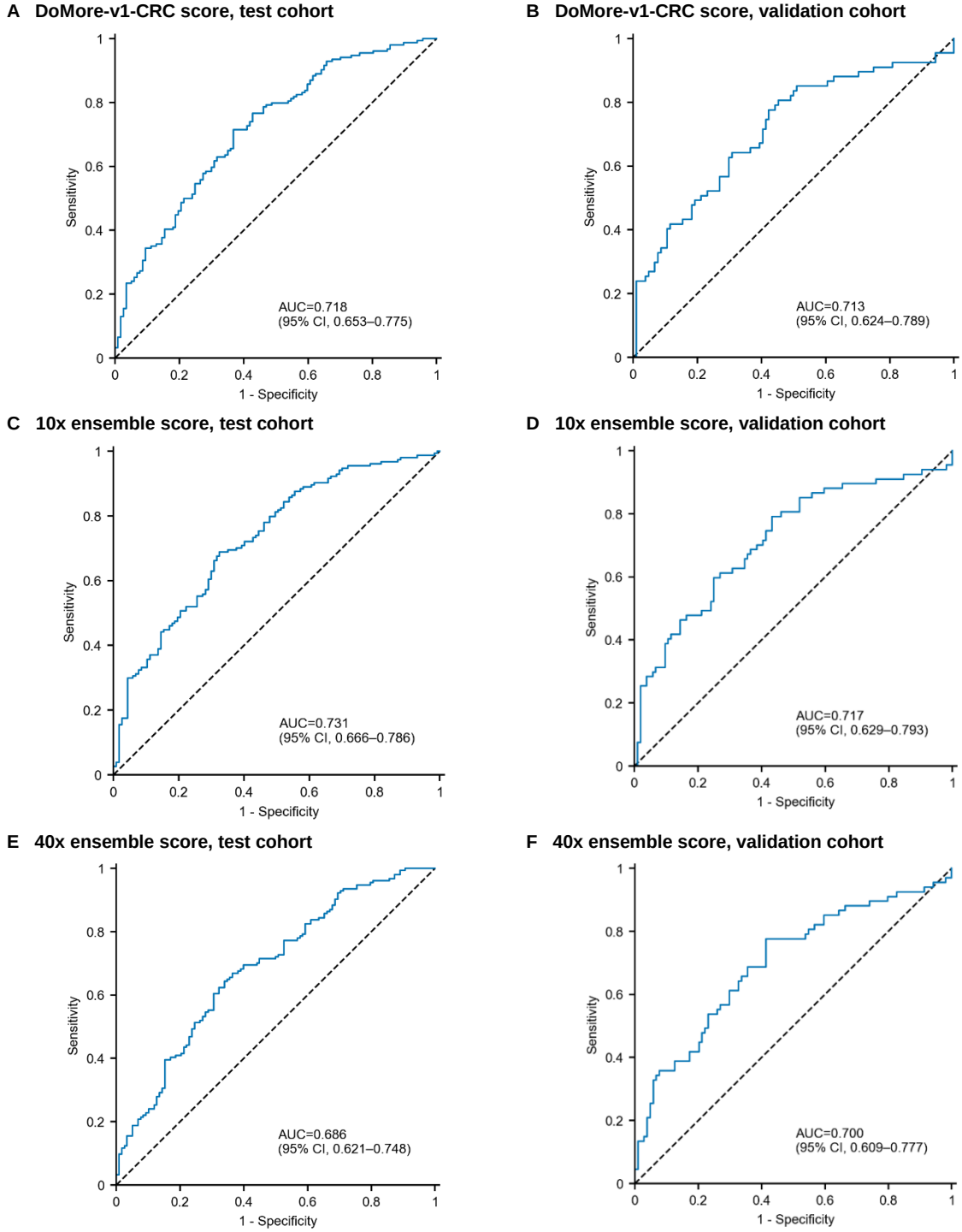
For each DoMore v1 40x individual model, the score was dichotomised using the same threshold as used for the DoMore v1 40x ensemble model.

Figure S5: Cancer-specific survival against DoMore-v1-CRC score evaluated on Aperio AT2 slide images



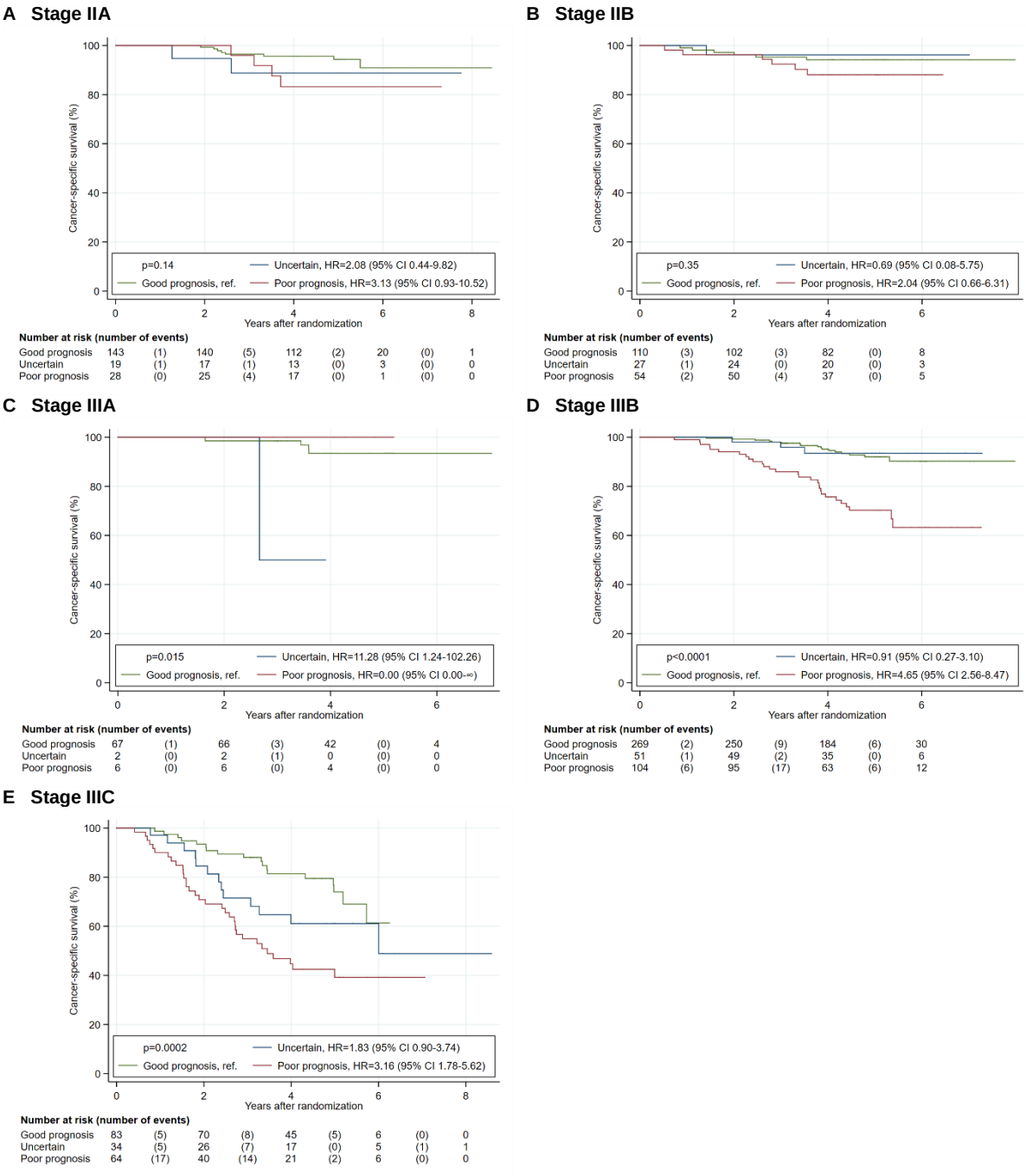
The probability of cancer-specific survival was estimated for scores 0, 0·01, and so on up to 1, as the proportion of cancer-specific deaths among the 20 patients with nearest score, and the corresponding 95% confidence intervals (CIs) were computed as the bias-corrected and accelerated bootstrap CIs. The estimated probabilities were then filtered by a 101 elements wide Gaussian kernel with standard deviation 0·1, as was also each of the CI limits.

Figure S6: Receiver operating characteristic (ROC) plots of the DoMore-v1-CRC score and its constituents evaluated on Aperio AT2 slide images from patients with distinct outcome



The associated areas under the curves (AUCs) are given with 95% confidence intervals (CIs). Distinct outcome in the test cohort was defined as in the training cohort, and similarly for the validation cohort; aged less than 85 years at randomisation and either more than 6 years follow-up after randomisation without record of recurrence or cancer-specific death, or suffered cancer-specific death between 100 days (inclusive) and 2·5 years (exclusive) after randomisation. 120 patients in the test cohort had good outcome and 157 had poor outcome, while 105 patients in the validation cohort had good outcome and 67 had poor outcome.

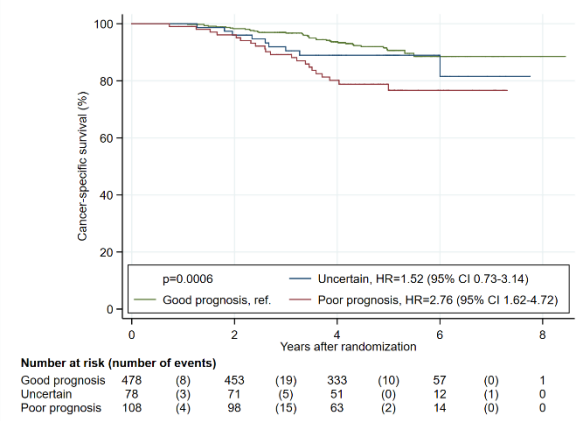
Figure S7: Kaplan-Meier analysis of cancer-specific survival by DoMore-v1-CRC class evaluated on Aperio AT2 slide images in each substage of the validation cohort



The substages were defined with respect to pTNM stage; stage IIA was T3, N0, M0, stage IIB was T4, N0, M0, stage IIIA was T1, N1, M0 or T2, N1, M0, stage IIIB was T3, N1, M0 or T4, N1, M0, and stage IIIC was any T, N2, M0.

Figure S8: Kaplan-Meier analysis of cancer-specific survival by DoMore-v1-CRC class evaluated on Aperio AT2 slide images in pT stages of the validation cohort

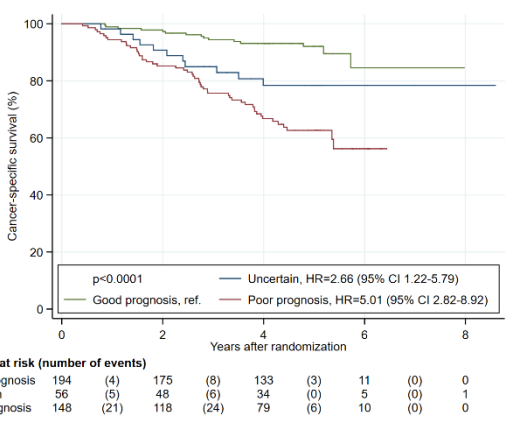
A pT1-3



Number at risk (number of events)

Good prognosis	478	(8)	453	(19)	333	(10)	57	(0)	1
Uncertain	78	(3)	71	(5)	51	(0)	12	(1)	0
Poor prognosis	108	(4)	98	(15)	63	(2)	14	(0)	0

B pT4

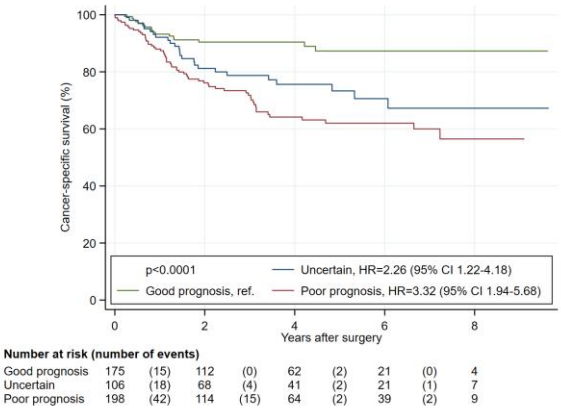


Number at risk (number of events)

Good prognosis	194	(4)	175	(8)	133	(3)	11	(0)	0
Uncertain	56	(5)	48	(6)	34	(0)	5	(0)	1
Poor prognosis	148	(21)	118	(24)	79	(6)	10	(0)	0

Figure S9: Kaplan-Meier analysis of cancer-specific survival by DoMore-v1-CRC class evaluated on Aperio AT2 slide images in patients with moderately differentiated tumours, i.e. histological grade 2

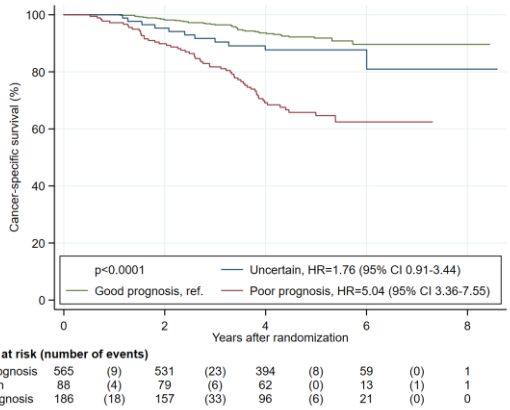
A Test cohort



Number at risk (number of events)

Good prognosis	175	(15)	112	(0)	62	(2)	21	(0)	4
Uncertain	106	(18)	68	(4)	41	(2)	21	(1)	7
Poor prognosis	198	(42)	114	(15)	64	(2)	39	(2)	9

B Validation cohort

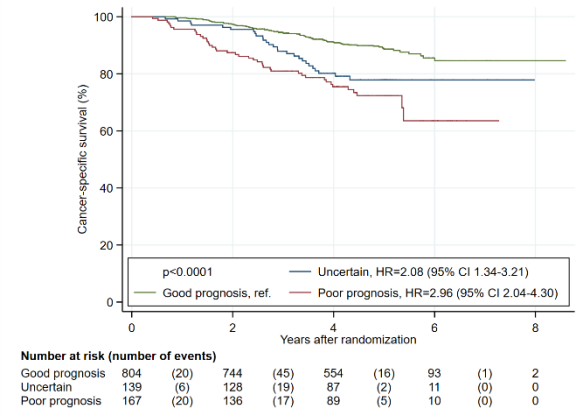


Number at risk (number of events)

Good prognosis	565	(9)	531	(23)	394	(8)	59	(0)	1
Uncertain	88	(4)	79	(6)	62	(0)	13	(1)	1
Poor prognosis	186	(18)	157	(33)	96	(6)	21	(0)	0

Figure S10: Kaplan-Meier analysis of cancer-specific survival by the Inception v3 classifier and its constituents evaluated on Aperio AT2 slide images in the validation cohort

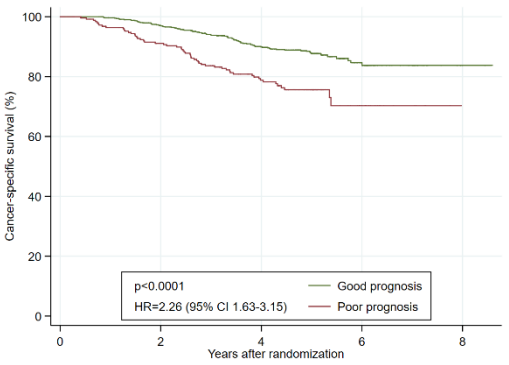
A Inception v3 classifier (secondary analysis)



Number at risk (number of events)

Good prognosis	804	(20)	744	(45)	554	(16)	93	(1)	2
Uncertain	139	(6)	128	(19)	87	(2)	11	(0)	0
Poor prognosis	167	(20)	136	(17)	89	(5)	10	(0)	0

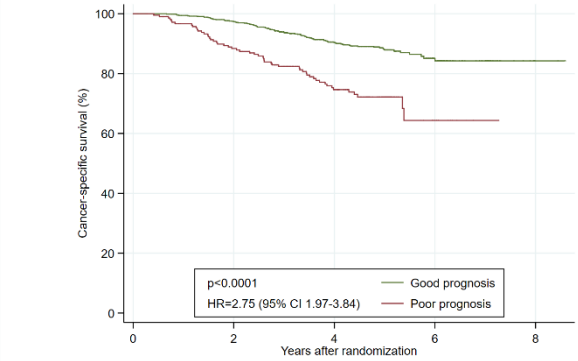
B 10x ensemble classifier (secondary analysis)



Number at risk (number of events)

Good prognosis	853	(24)	786	(53)	579	(16)	95	(1)	2
Poor prognosis	257	(22)	222	(28)	151	(7)	19	(0)	0

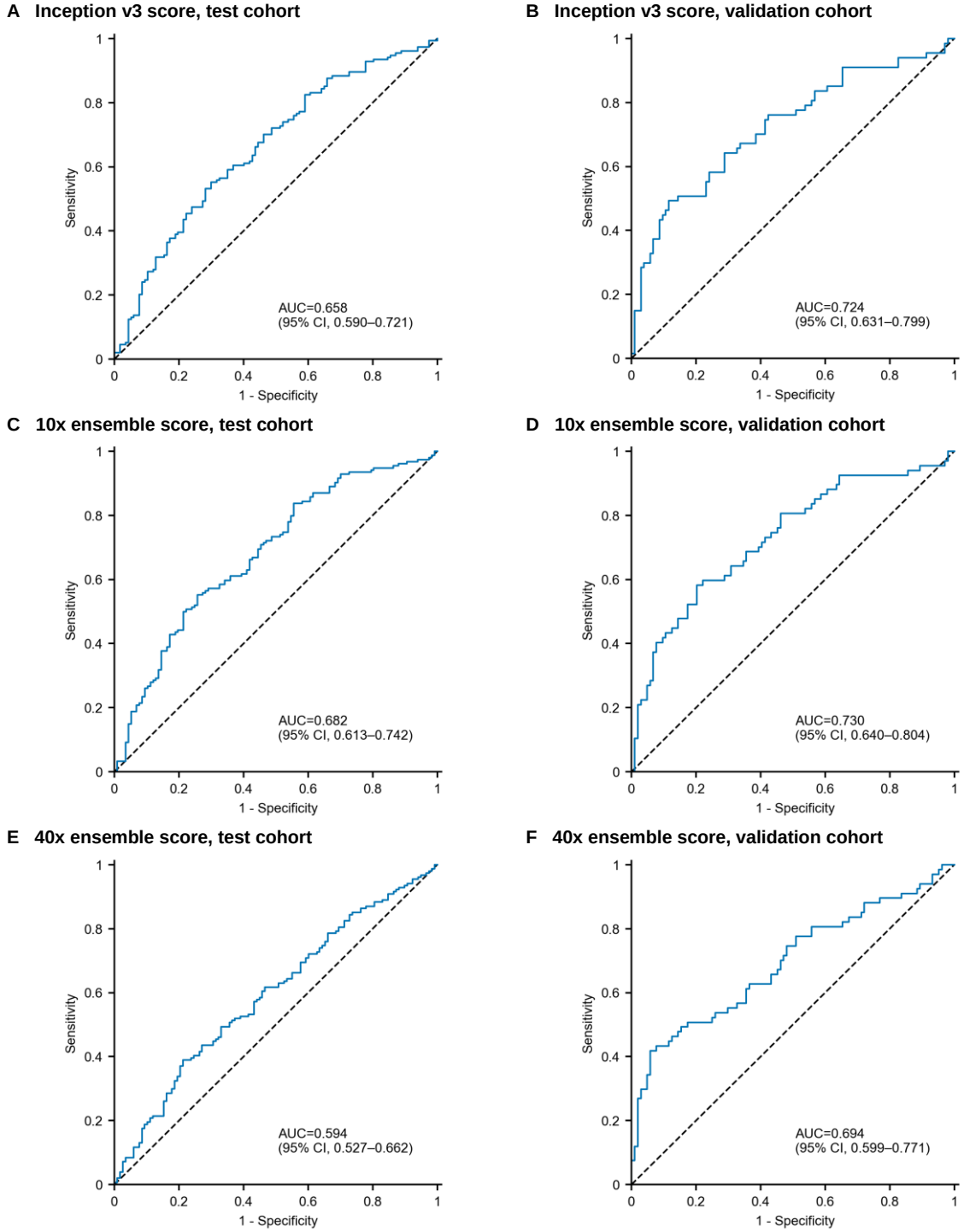
C 40x ensemble classifier (secondary analysis)



Number at risk (number of events)

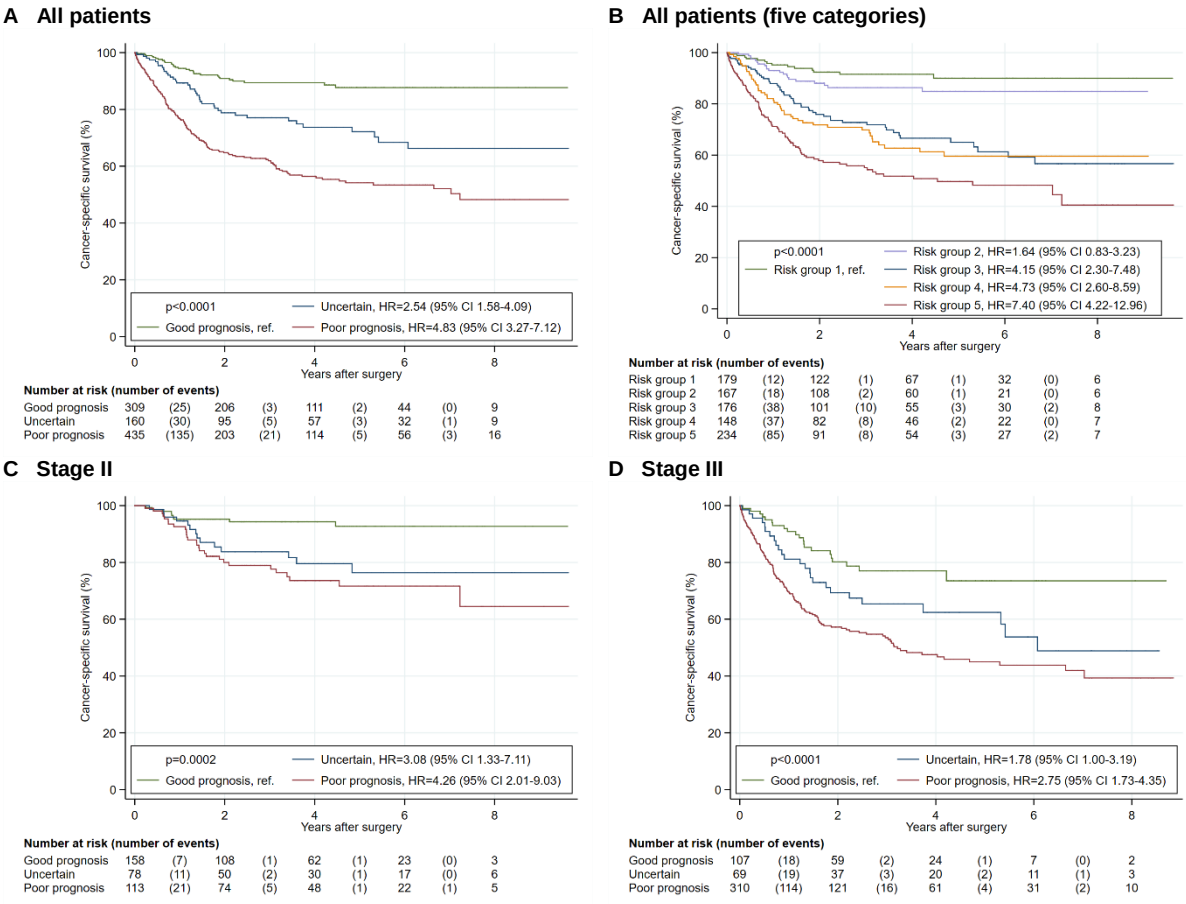
Good prognosis	896	(22)	832	(56)	617	(18)	102	(1)	2
Poor prognosis	217	(24)	179	(25)	115	(5)	12	(0)	0

Figure S11: Receiver operating characteristic (ROC) plots of the Inception v3 score and its constituents evaluated on Aperio AT2 slide images from patients with distinct outcome



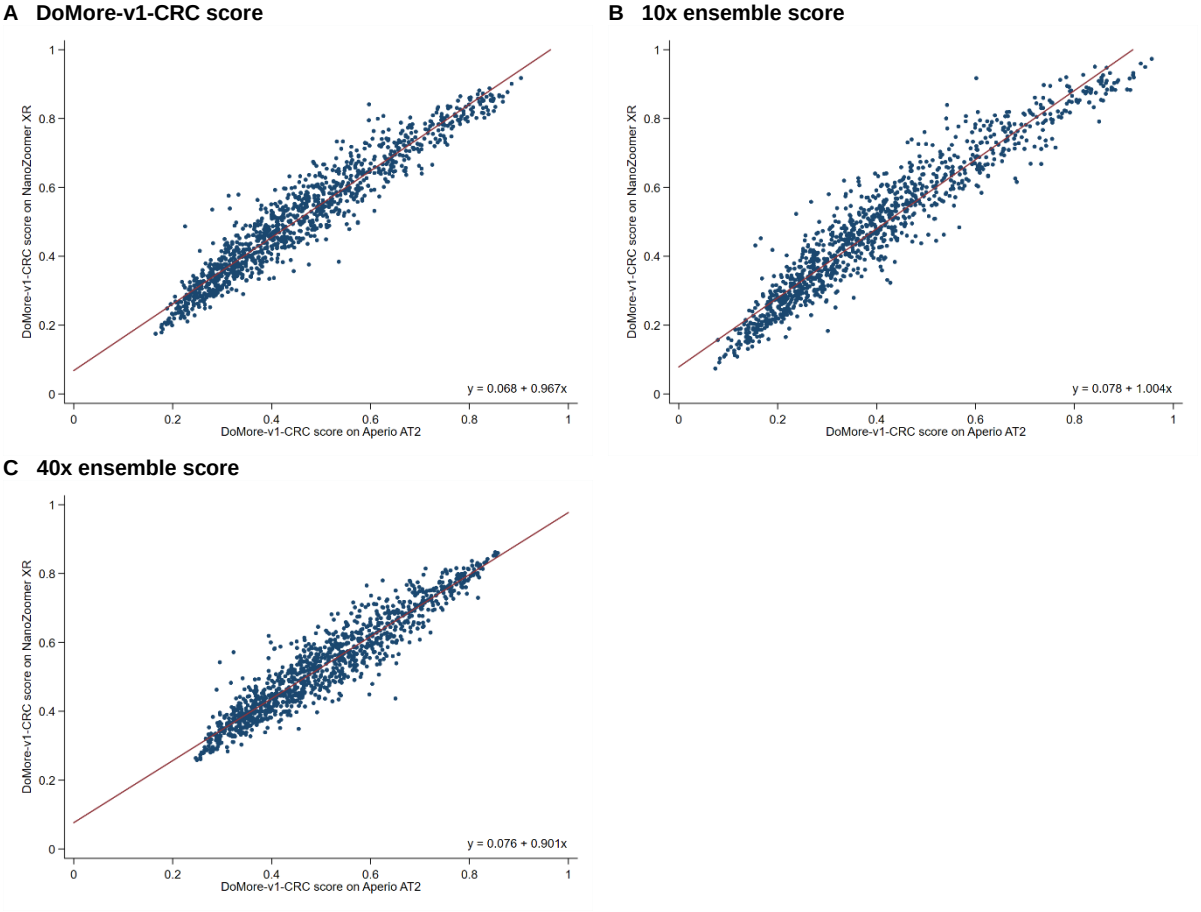
The associated areas under the curves (AUCs) are given with 95% confidence intervals (CIs). Distinct outcome in the test cohort was defined as in the training cohort, and similarly for the validation cohort; aged less than 85 years at randomisation and either more than 6 years follow-up after randomisation without record of recurrence or cancer-specific death, or suffered cancer-specific death between 100 days (inclusive) and 2·5 years (exclusive) after randomisation. 120 patients in the test cohort had good outcome and 157 had poor outcome, while 105 patients in the validation cohort had good outcome and 67 had poor outcome.

Figure S12: Kaplan-Meier analysis of cancer-specific survival by DoMore-v1-CRC class evaluated on Aperio AT2 slide images in the test cohort



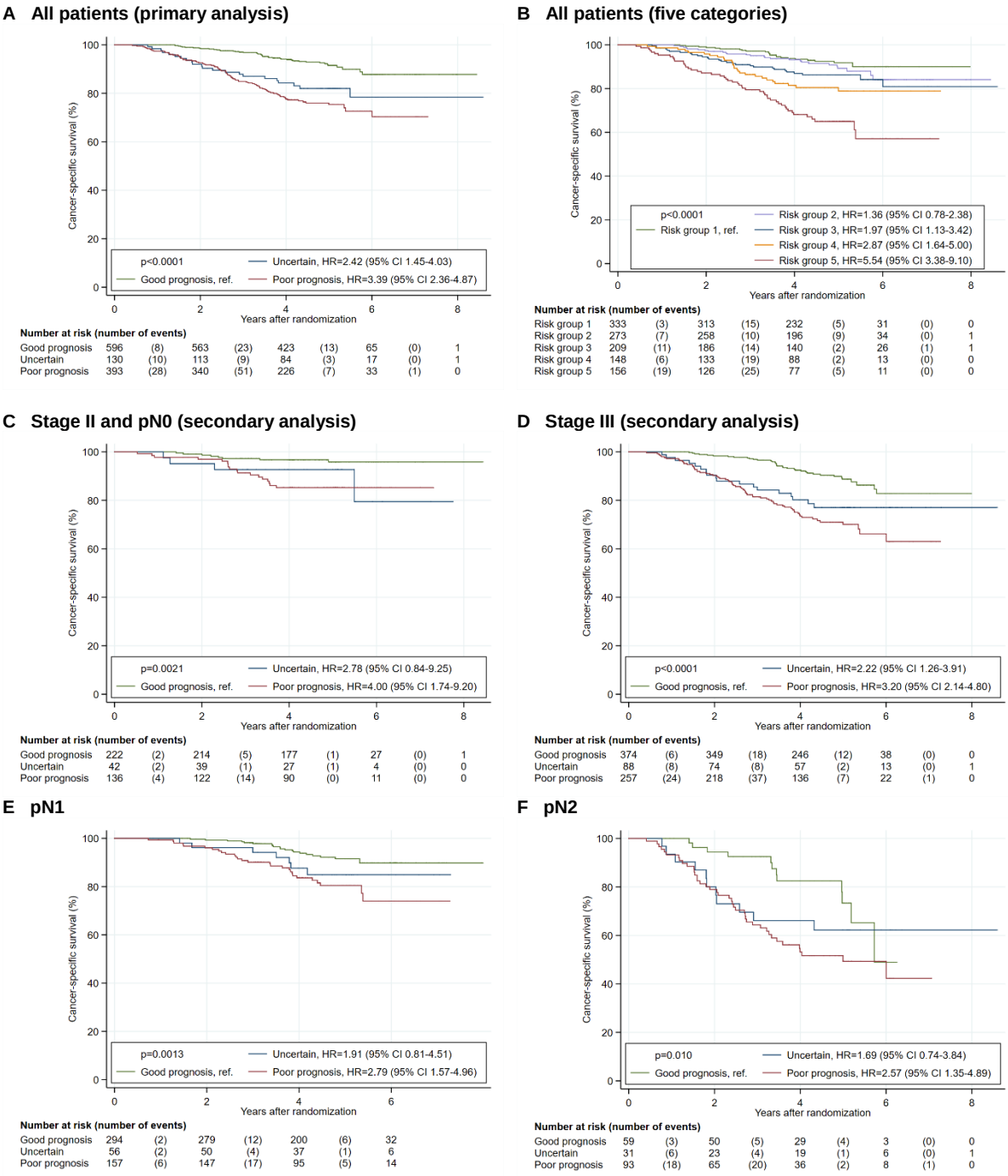
The pre-defined DoMore-v1-CRC classifier was evaluated in Panels A, C, and D. The DoMore-v1-CRC classifier variant with five categories was evaluated in Panel B.

Figure S13: Scatter plot of the DoMore-v1-CRC score and its constituents evaluated on NanoZoomer XR vs. Aperio AT2 slide images in the validation cohort



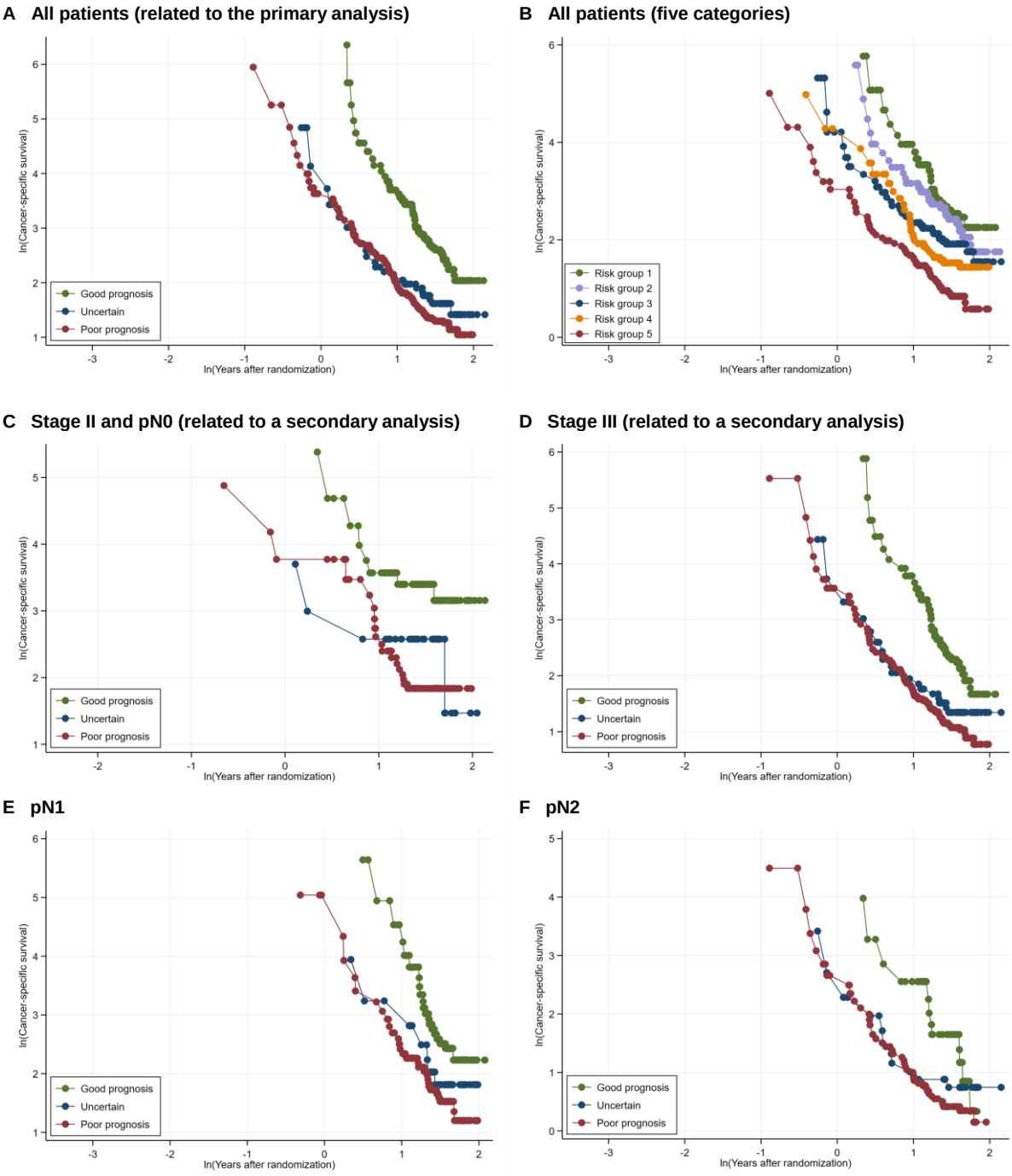
Pearson correlation coefficient was 0·956 (95% CI, 0·951-0·961; p<0·0001) for the DoMore-v1-CRC score, 0·954 (95% CI, 0·948-0·959; p<0·0001) for the 10x ensemble score of the DoMore v1 network, and 0·944 (95% CI, 0·937-0·950; p<0·0001) for the 40x ensemble score of the DoMore v1 network.

Figure S14: Kaplan-Meier analysis of cancer-specific survival by DoMore-v1-CRC class evaluated on NanoZoomer XR slide images in the validation cohort



The pre-defined DoMore-v1-CRC classifier was evaluated in Panels A, C, D, E, and F. The DoMore-v1-CRC classifier variant with five categories was evaluated in Panel B.

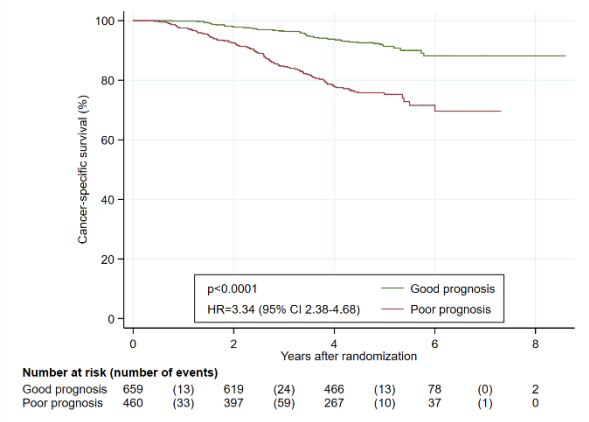
Figure S15: log-log plots of cancer-specific survival by DoMore-v1-CRC class evaluated on NanoZoomer XR slide images in the validation cohort (comparable to figure S10)



The pre-defined DoMore-v1-CRC classifier was evaluated in Panels A, C, D, E, and F. The DoMore-v1-CRC classifier variant with five categories was evaluated in Panel B.

Figure S16: Kaplan-Meier analysis of cancer-specific survival by the constituents of the DoMore-v1-CRC class evaluated on NanoZoomer XR slide images in the validation cohort

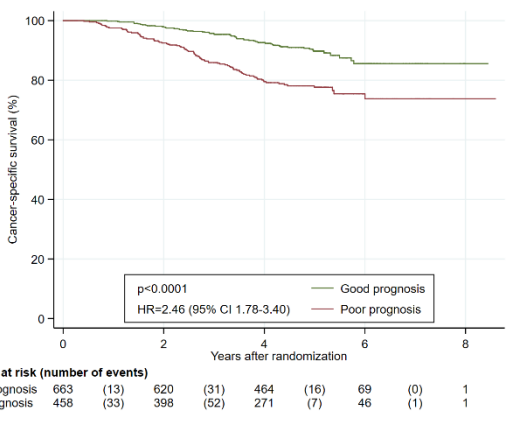
A 10x ensemble classifier (secondary analysis)



Number at risk (number of events)

Good prognosis	659	(13)	619	(24)	466	(13)	78	(0)	2
Poor prognosis	460	(33)	397	(59)	267	(10)	37	(1)	0

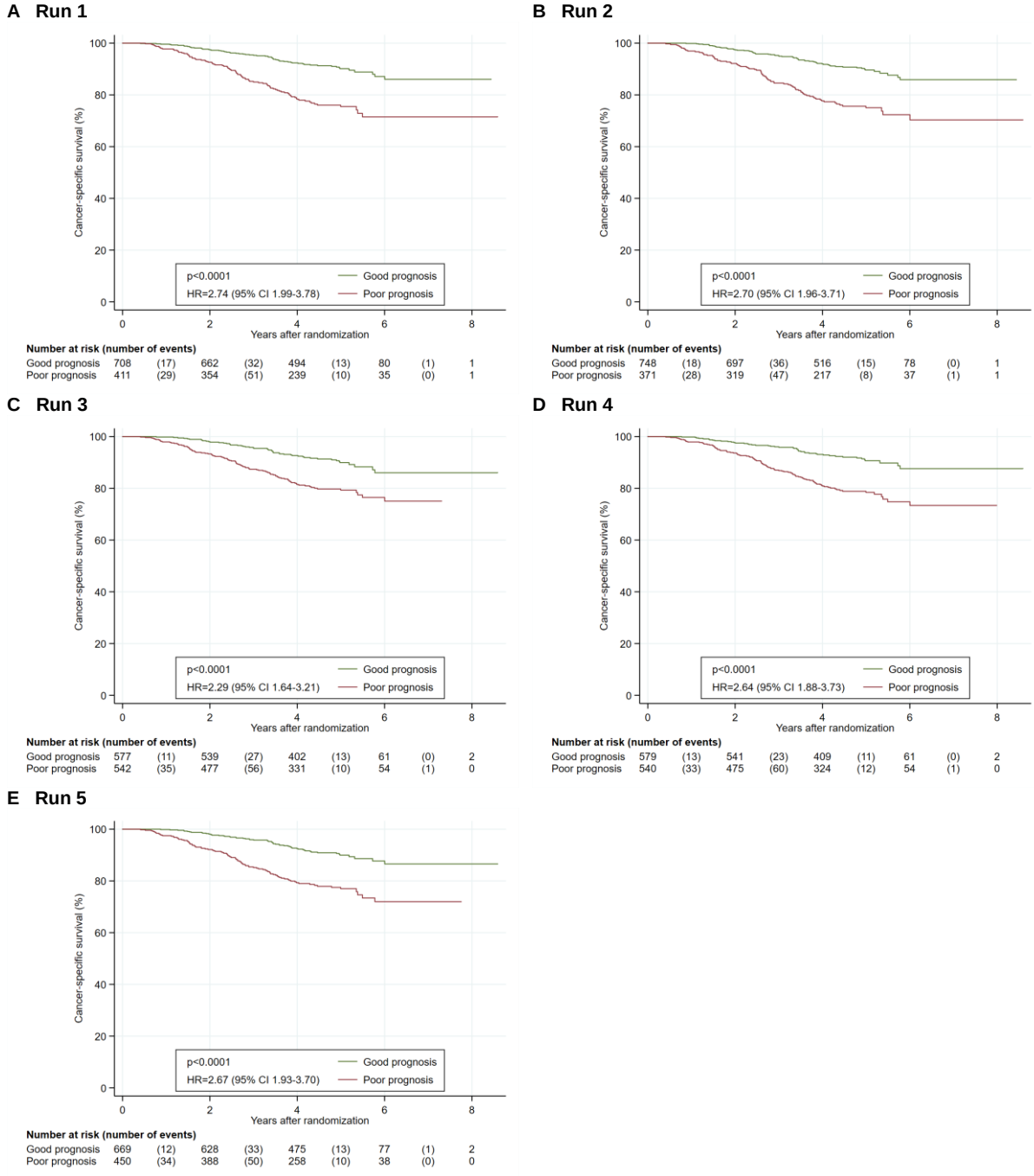
B 40x ensemble classifier (secondary analysis)



Number at risk (number of events)

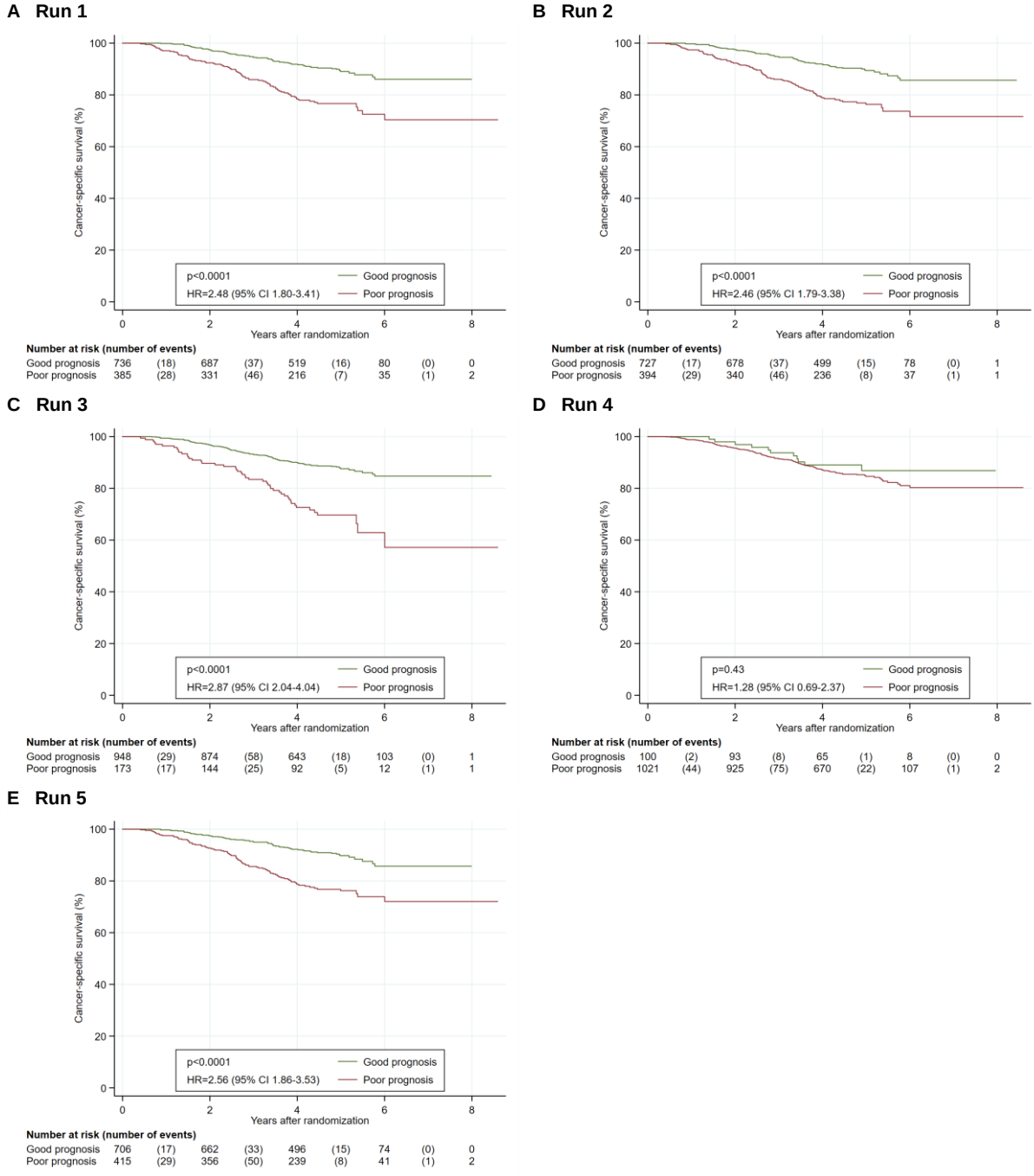
Good prognosis	663	(13)	620	(31)	464	(16)	69	(0)	1
Poor prognosis	458	(33)	398	(52)	271	(7)	46	(1)	1

Figure S17: Kaplan-Meier analysis of cancer-specific survival by the class of a DoMore v1 10x individual model evaluated on NanoZoomer XR slide images in the validation cohort



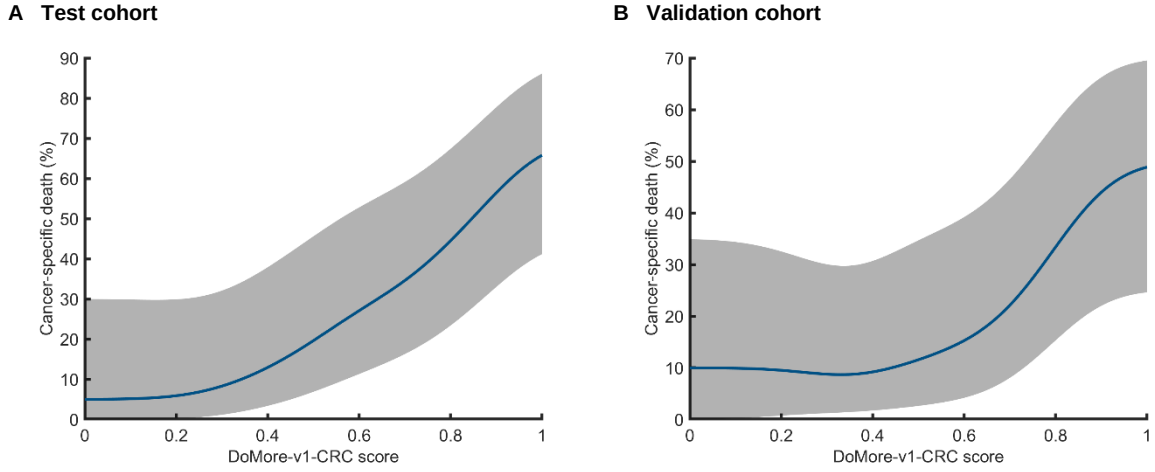
For each DoMore v1 10x individual model, the score was dichotomised using the same threshold as used for the DoMore v1 10x ensemble model.

Figure S18: Kaplan-Meier analysis of cancer-specific survival by the class of a DoMore v1 40x individual model evaluated on NanoZoomer XR slide images in the validation cohort



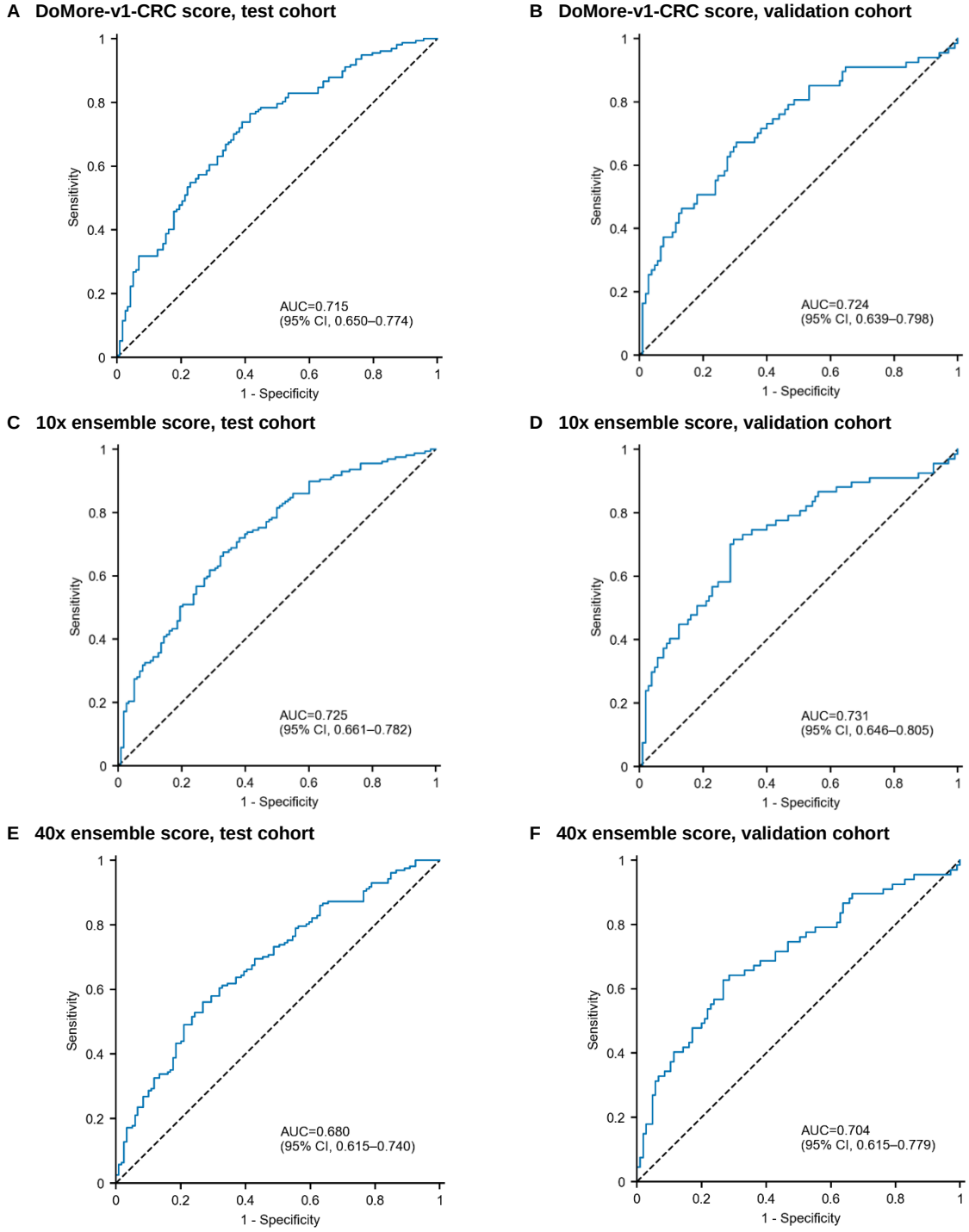
For each DoMore v1 40x individual model, the score was dichotomised using the same threshold as used for the DoMore v1 40x ensemble model.

Figure S19: Cancer-specific survival against DoMore-v1-CRC score evaluated on NanoZoomer XR slide images



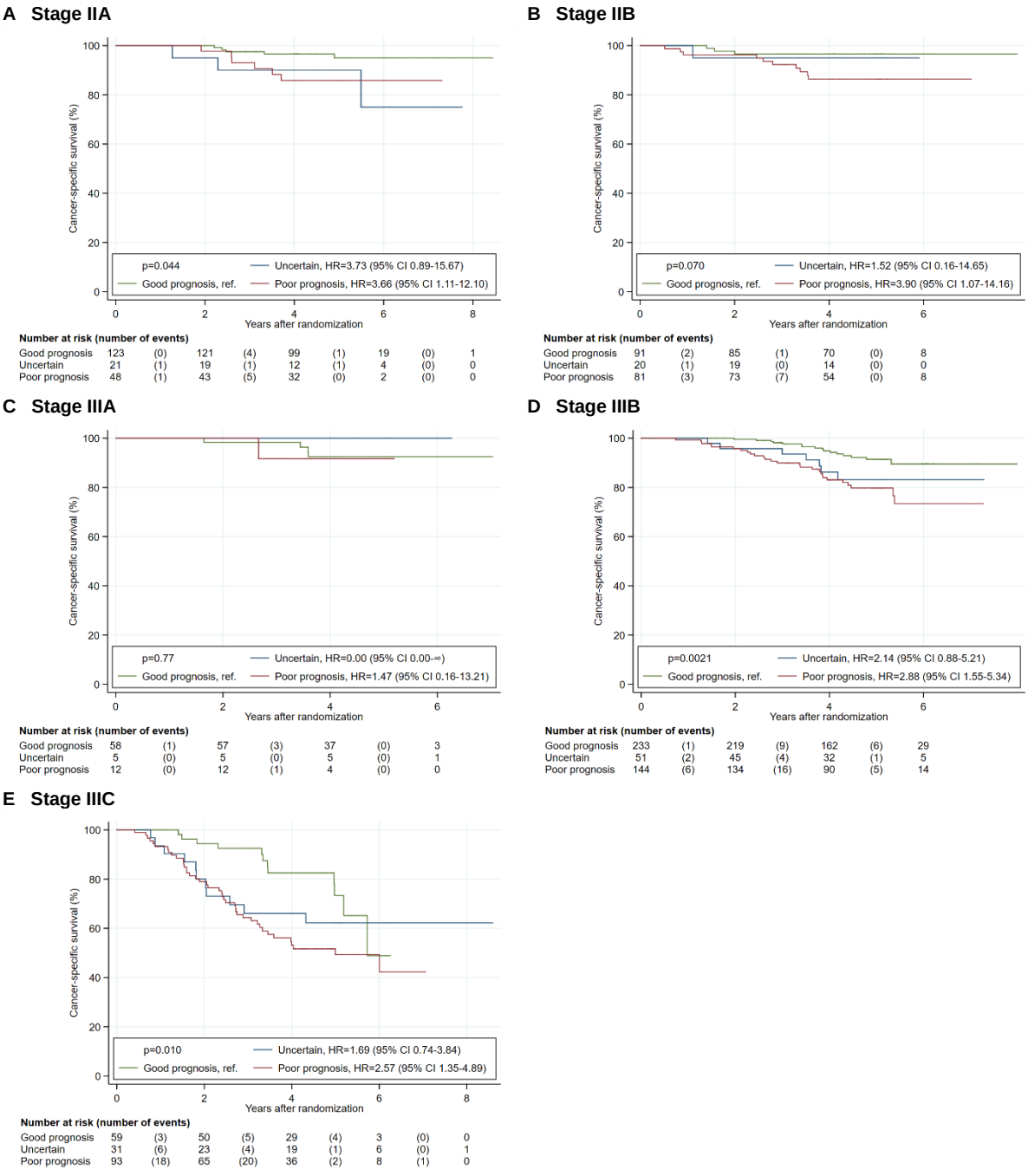
The probability of cancer-specific survival was estimated for scores 0, 0·01, and so on up to 1, as the proportion of cancer-specific deaths among the 20 patients with nearest score, and the corresponding 95% confidence intervals (CIs) were computed as the bias-corrected and accelerated bootstrap CIs. The estimated probabilities were then filtered by a 101 elements wide Gaussian kernel with standard deviation 0·1, as was also each of the CI limits.

Figure S20: Receiver operating characteristic (ROC) plots of the DoMore-v1-CRC score and its constituents evaluated on NanoZoomer XR slide images from patients with distinct outcome



The associated areas under the curves (AUCs) are given with 95% confidence intervals (CIs). Distinct outcome in the test cohort was defined as in the training cohort, and similarly for the validation cohort; aged less than 85 years at randomisation and either more than 6 years follow-up after randomisation without record of recurrence or cancer-specific death, or suffered cancer-specific death between 100 days (inclusive) and 2·5 years (exclusive) after randomisation. 120 patients in the test cohort had good outcome and 157 had poor outcome, while 105 patients in the validation cohort had good outcome and 67 had poor outcome.

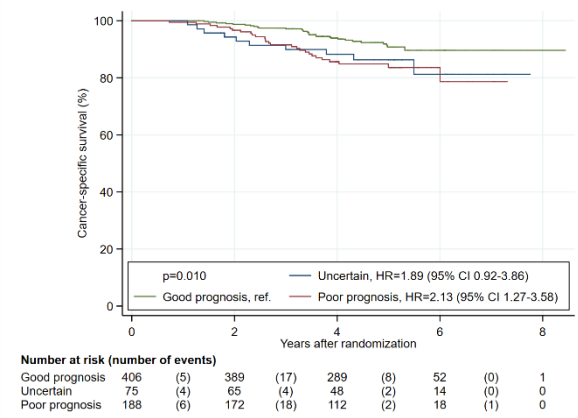
Figure S21: Kaplan-Meier analysis of cancer-specific survival by DoMore-v1-CRC class evaluated on NanoZoomer XR slide images in each substage of the validation cohort



The substages were defined with respect to pTNM stage; stage IIA was T3, N0, M0, stage IIB was T4, N0, M0, stage IIIA was T1, N1, M0 or T2, N1, M0, stage IIIB was T3, N1, M0 or T4, N1, M0, and stage IIIC was any T, N2, M0.

Figure S22: Kaplan-Meier analysis of cancer-specific survival by DoMore-v1-CRC class evaluated on NanoZoomer XR slide images in pT stages of the validation cohort

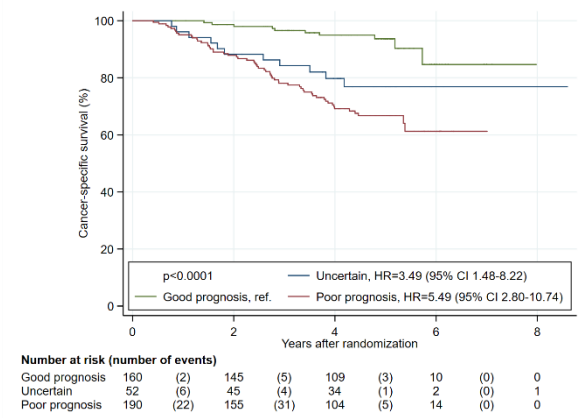
A pT1-3



Number at risk (number of events)

Good prognosis	406	(5)	389	(17)	289	(8)	52	(0)	1
Uncertain	75	(4)	65	(4)	48	(2)	14	(0)	0
Poor prognosis	188	(6)	172	(18)	112	(2)	18	(1)	0

B pT4

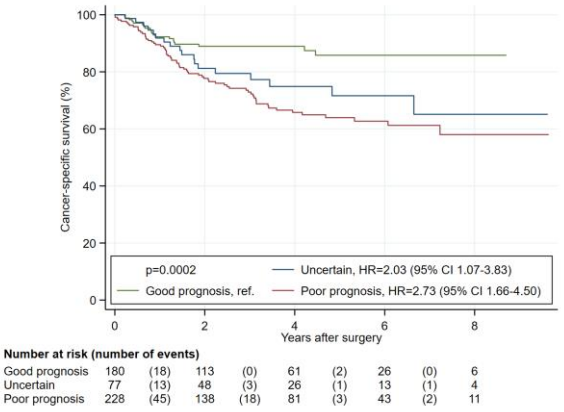


Number at risk (number of events)

Good prognosis	160	(2)	145	(5)	109	(3)	10	(0)	0
Uncertain	52	(6)	45	(4)	34	(1)	2	(0)	1
Poor prognosis	190	(22)	155	(31)	104	(5)	14	(0)	0

Figure S23: Kaplan-Meier analysis of cancer-specific survival by DoMore-v1-CRC class evaluated on NanoZoomer XR slide images in patients with moderately differentiated tumours, i.e. histological grade 2

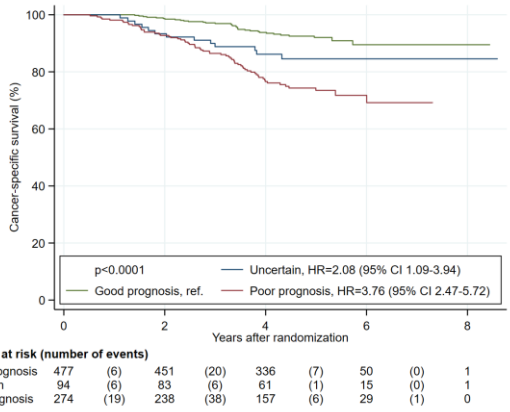
A Test cohort



Number at risk (number of events)

Good prognosis	180	(18)	113	(0)	61	(2)	26	(0)	6
Uncertain	77	(13)	48	(3)	26	(1)	13	(1)	4
Poor prognosis	228	(45)	138	(18)	81	(3)	43	(2)	11

B Validation cohort

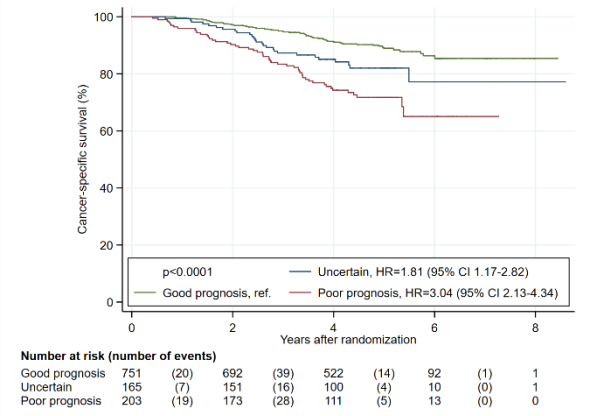


Number at risk (number of events)

Good prognosis	477	(6)	451	(20)	336	(7)	50	(0)	1
Uncertain	94	(6)	83	(6)	61	(1)	15	(0)	1
Poor prognosis	274	(19)	238	(38)	157	(6)	29	(1)	0

Figure S24: Kaplan-Meier analysis of cancer-specific survival by the Inception v3 classifier and its constituents evaluated on NanoZoomer XR slide images in the validation cohort

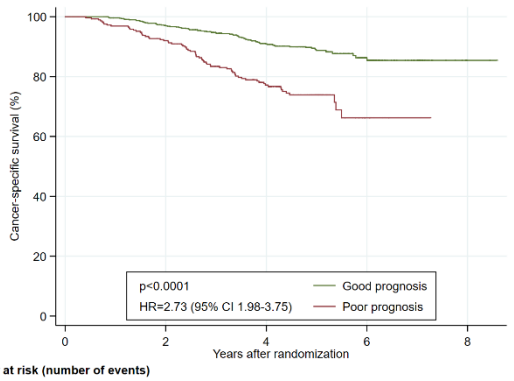
A Inception v3 classifier (secondary analysis)



Number at risk (number of events)

Good prognosis	751	(20)	692	(39)	522	(14)	92	(1)	1
Uncertain	165	(7)	151	(16)	100	(4)	10	(0)	1
Poor prognosis	203	(19)	173	(28)	111	(5)	13	(0)	0

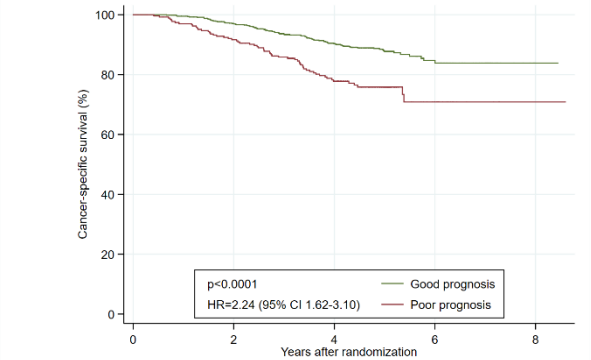
B 10x ensemble classifier (secondary analysis)



Number at risk (number of events)

Good prognosis	822	(23)	755	(44)	565	(14)	99	(1)	2
Poor prognosis	297	(23)	261	(39)	168	(9)	16	(0)	0

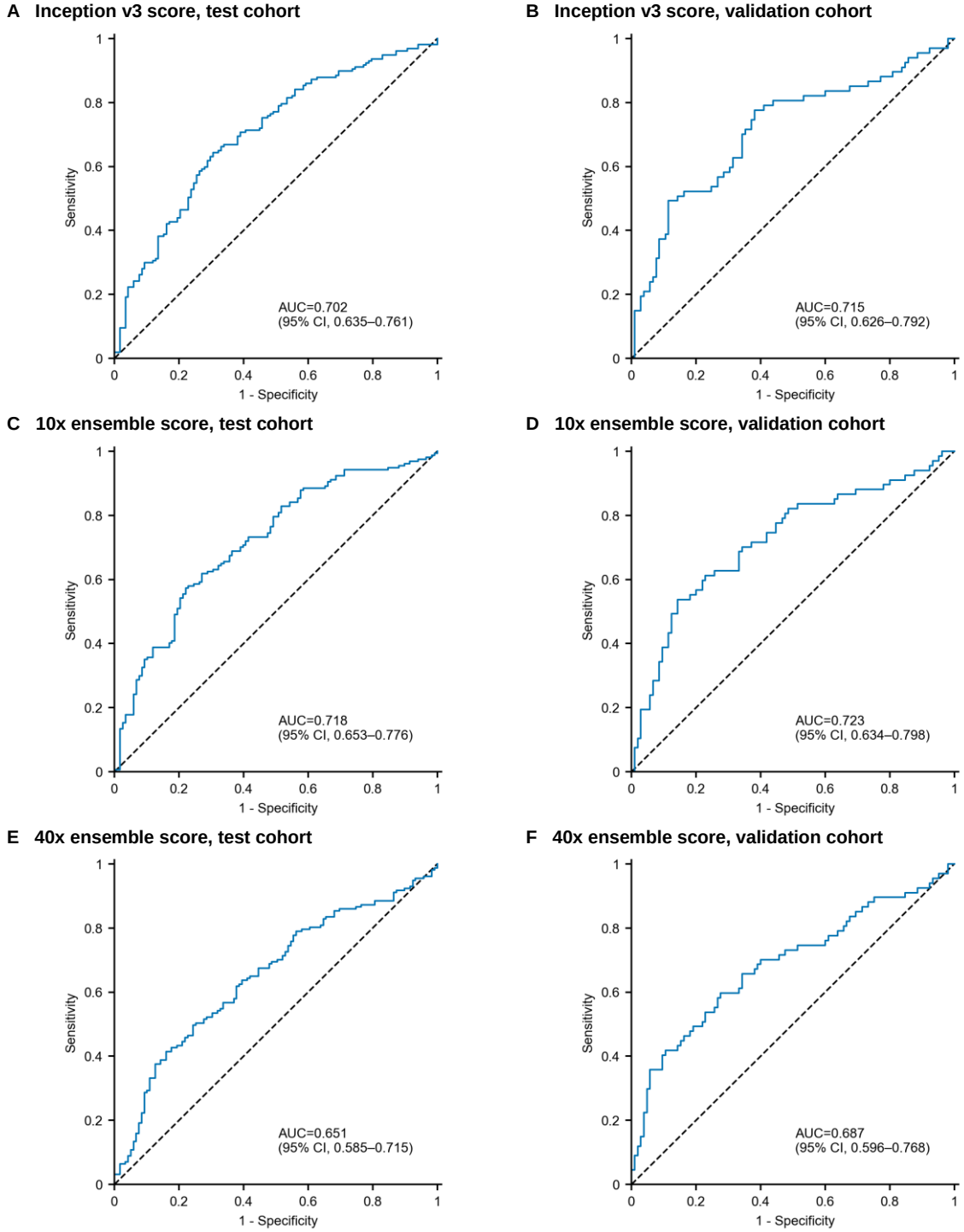
C 40x ensemble classifier (secondary analysis)



Number at risk (number of events)

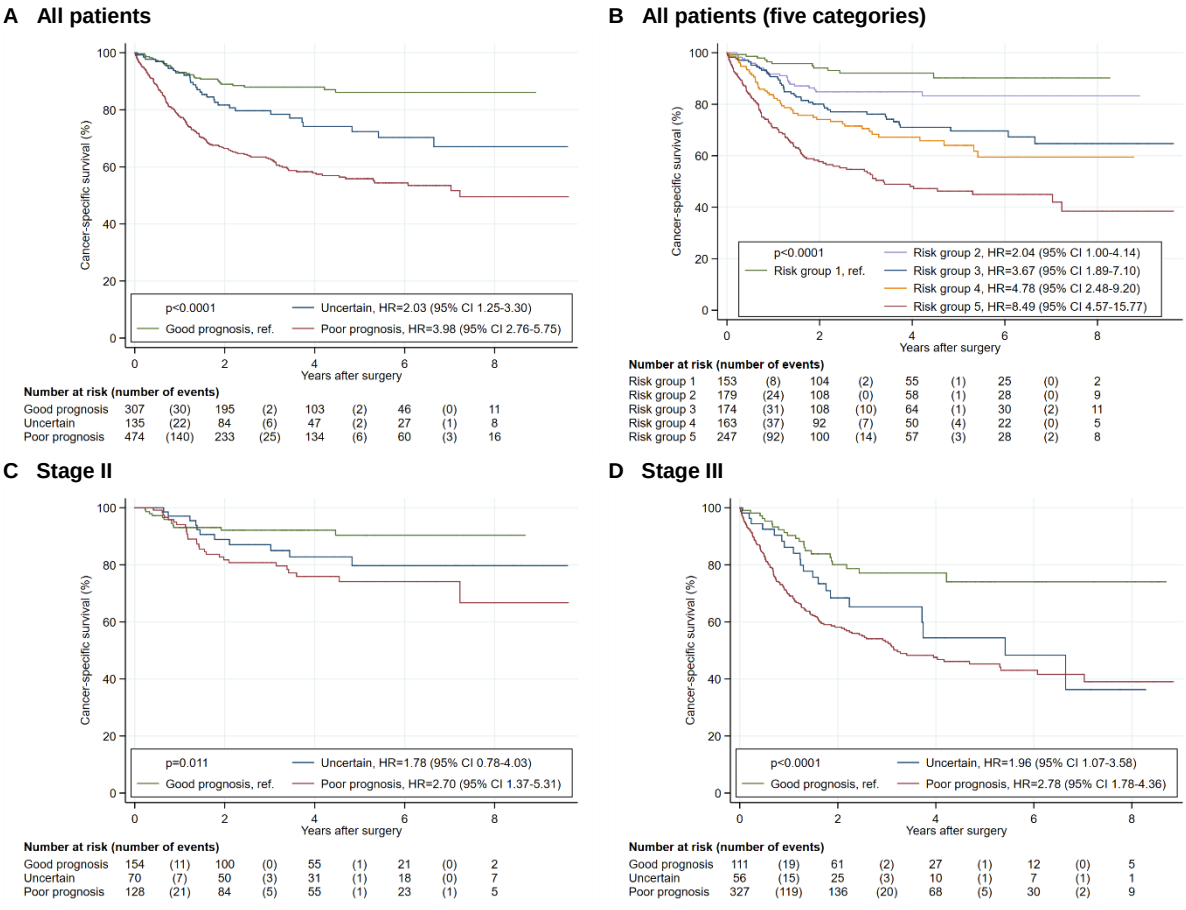
Good prognosis	845	(24)	780	(50)	579	(18)	95	(1)	1
Poor prognosis	276	(22)	238	(33)	156	(5)	20	(0)	1

Figure S25: Receiver operating characteristic (ROC) plots of the Inception v3 score and its constituents evaluated on NanoZoomer XR slide images from patients with distinct outcome



The associated areas under the curves (AUCs) are given with 95% confidence intervals (CIs). Distinct outcome in the test cohort was defined as in the training cohort, and similarly for the validation cohort; aged less than 85 years at randomisation and either more than 6 years follow-up after randomisation without record of recurrence or cancer-specific death, or suffered cancer-specific death between 100 days (inclusive) and 2·5 years (exclusive) after randomisation. 120 patients in the test cohort had good outcome and 157 had poor outcome, while 105 patients in the validation cohort had good outcome and 67 had poor outcome.

Figure S26: Kaplan-Meier analysis of cancer-specific survival by DoMore-v1-CRC class evaluated on NanoZoomer XR slide images in the test cohort



The pre-defined DoMore-v1-CRC classifier was evaluated in Panels A, C, and D. The DoMore-v1-CRC classifier variant with five categories was evaluated in Panel B.

External evaluation of a deep learning model for prediction of cancer-specific survival from colorectal cancer tissue sections

1 Status at last amended

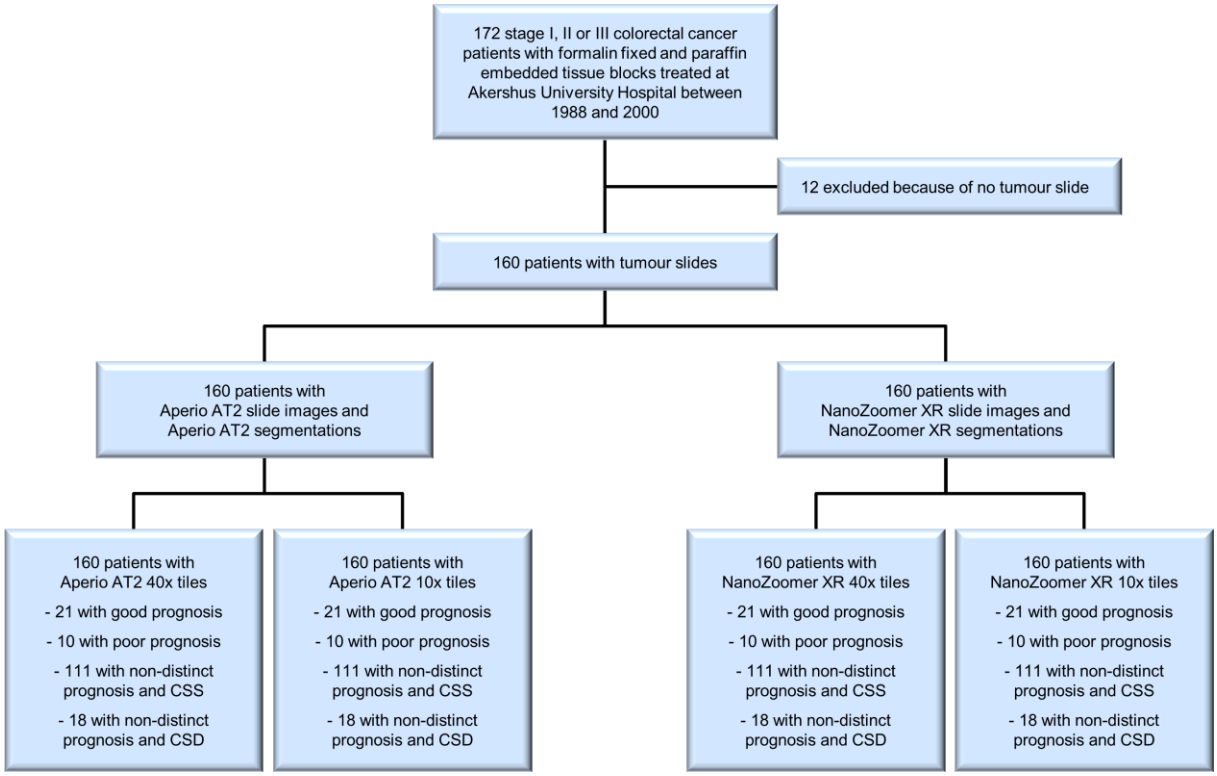
This protocol was last modified on 18th of March 2019, prior to all investigations that could reveal associations between slides and clinical outcome in the QUASAR 2 cohort. At that time the slides in the QUASAR 2 cohort had been scanned, segmented and tiled blinded to the clinical outcome, i.e. the recorded cancer-specific survival.

2 Training cohorts

Four training cohorts were utilised in this study. These were the Ahus cohort, the Aker cohort, the Gloucester cohort and the VICTOR cohort that are described in the following subsections. Patients in the training cohorts were labelled as distinct or non-distinct prognosis depending on age at surgery and follow-up data. The distinct prognosis patients are comprised of patients defined as good prognosis and patients defined as poor prognosis. A patient was defined as good prognosis if aged less than 85 years at surgery, had more than 6 years follow-up after surgery, had no record of cancer-specific death and no record of recurrence. The availability of recurrence data varied between the cohorts and was particularly limited for the Gloucester cohort. For the Ahus cohort, good prognosis patients were required to have no record of metastasis (records of local recurrences were not available), while no record of local or metastatic recurrence were required for Aker, Gloucester and VICTOR patients. A patient was defined as poor prognosis if aged less than 85 years at surgery and suffered cancer-specific death between 100 days (inclusive) and 2.5 years (exclusive) after surgery. Patients not satisfying the criteria for either good or poor prognosis were defined as non-distinct prognosis.

2.1 Ahus cohort

From a consecutive series of 219 patients with colonic adenocarcinoma treated between 1988 and 2000 at Akershus University Hospital, Norway[1], 172 patients had stage I, II or III disease and accessible formalin-fixed, paraffin-embedded (FFPE) tissue blocks. A 3 µm section of each FFPE tumour tissue block was stained with

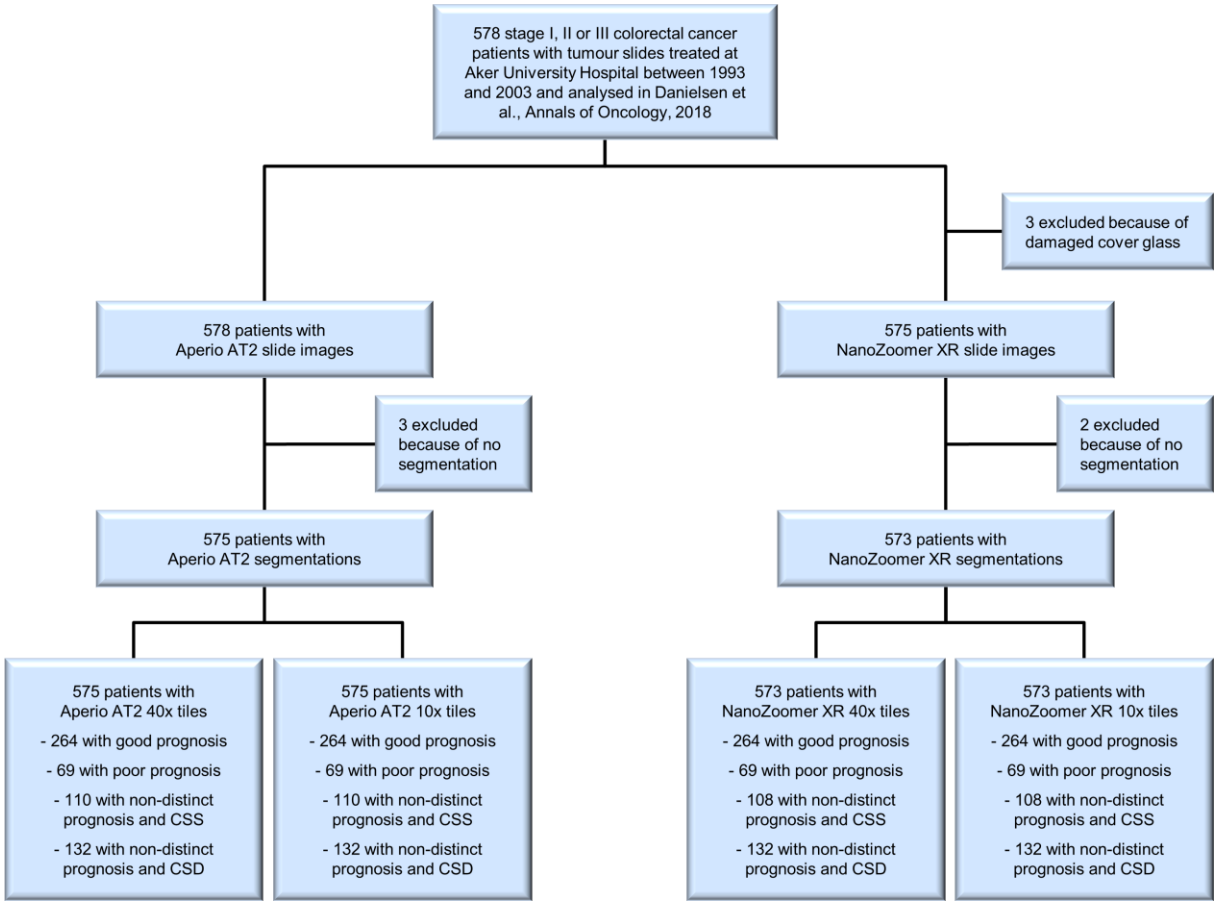


Protocol Fig. 1 | A diagram specifying inclusions and exclusions of patients, slides and slide images from the Ahus cohort, and the prognosis of the included patients. CSS, cancer-specific survival; CSD, cancer-specific death.

haematoxylin and eosin (H&E) and prepared as tissue slides by laboratory personnel at the Institute for Cancer Genetics and Informatics (ICGI), Oslo University Hospital, Norway. A pathologist ascertained whether there was tumour in each tissue section; the 12 patients without tumour slide were excluded (Protocol Fig. 1). The tumour tissue slides were scanned using two scanners, an Aperio AT2 (Leica Biosystems, Germany) and a NanoZoomer XR (Hamamatsu Photonics, Japan). The scans were read using the Python interface (version 1.1.1) of OpenSlide 3.4.1, available at https://openslide.org/. An automatic segmentation method (see section 3) was applied to identify tumour in the 320 slide images, and each slide image were partitioned into multiple non-overlapping regions called *tiles* using two resolution referred to as 40x and 10x (see section 4). The 160 included patients with tiles within the tumour segmentation were defined as the Ahus cohort; Protocol Fig. 1 specifies the prognosis of these patients (see definition of distinct and non-distinct prognosis in section 2 above).

2.2 Aker cohort

One slide from each of the 578 stage I, II or III patients treated between 1993 and 2003 for primary colorectal cancer at Aker University Hospital, Norway, and analysed by Danielsen and colleagues[2] were processed in the same manner as for the Ahus cohort. Three slides had damaged cover glass and could therefore not be scanned by the NanoZoomer XR scanner, and the automatic segmentation method identified no tumour for three Aperio AT2 slide images and two NanoZoomer XR slide images; the other patients comprised the Aker cohort (Protocol Fig. 2).

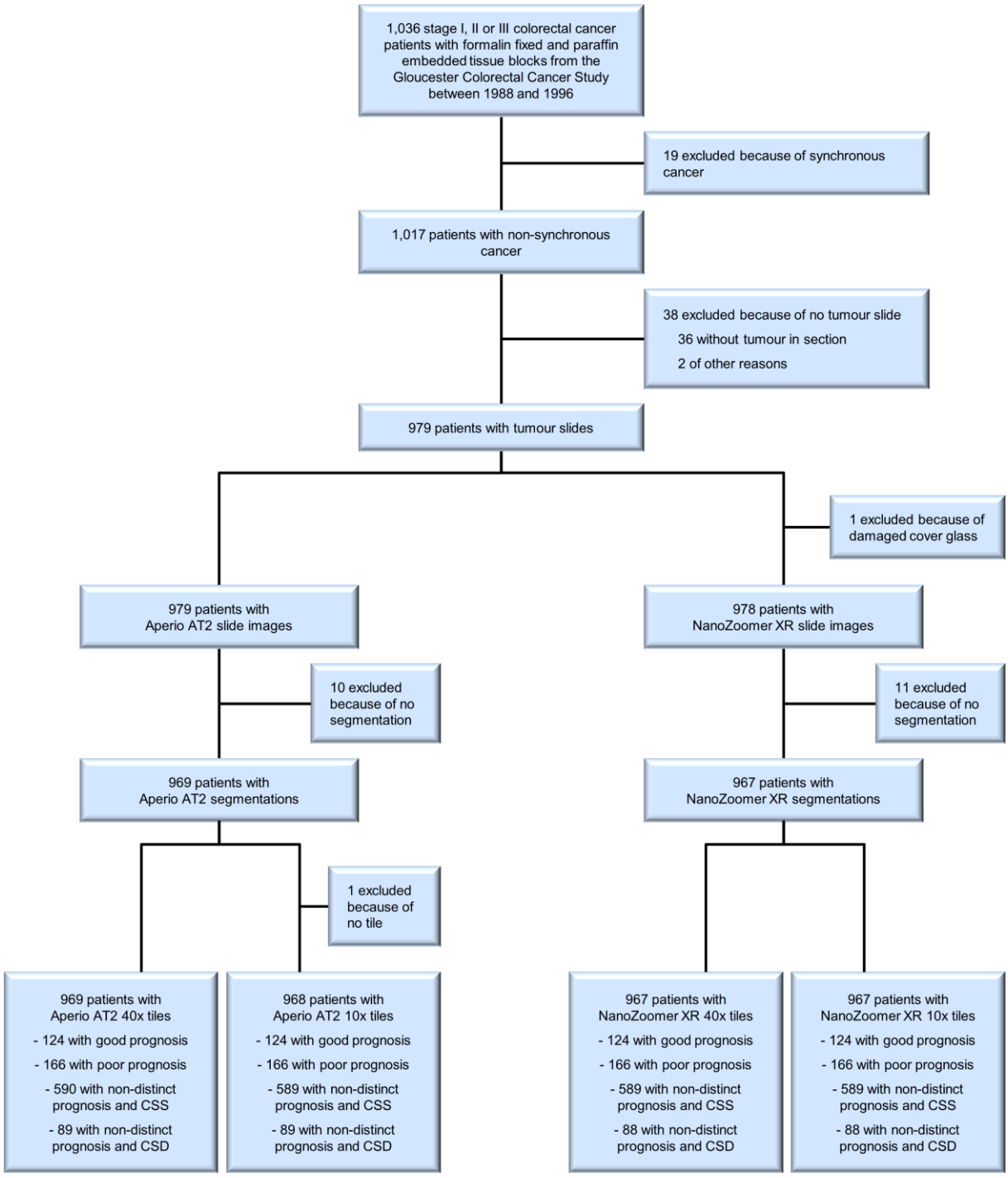


Protocol Fig. 2 | A diagram specifying inclusions and exclusions of patients, slides and slide images from the Aker cohort, and the prognosis of the included patients. CSS, cancer-specific survival; CSD, cancer-specific death.

2.3 Gloucester cohort

The Gloucester Colorectal Cancer Study recruited 1,036 patients between 1988 and 1996, of which 19 were excluded because of synchronous cancer (Protocol Fig. 3)[3,4]. The remaining 1,017 patients were processed in the

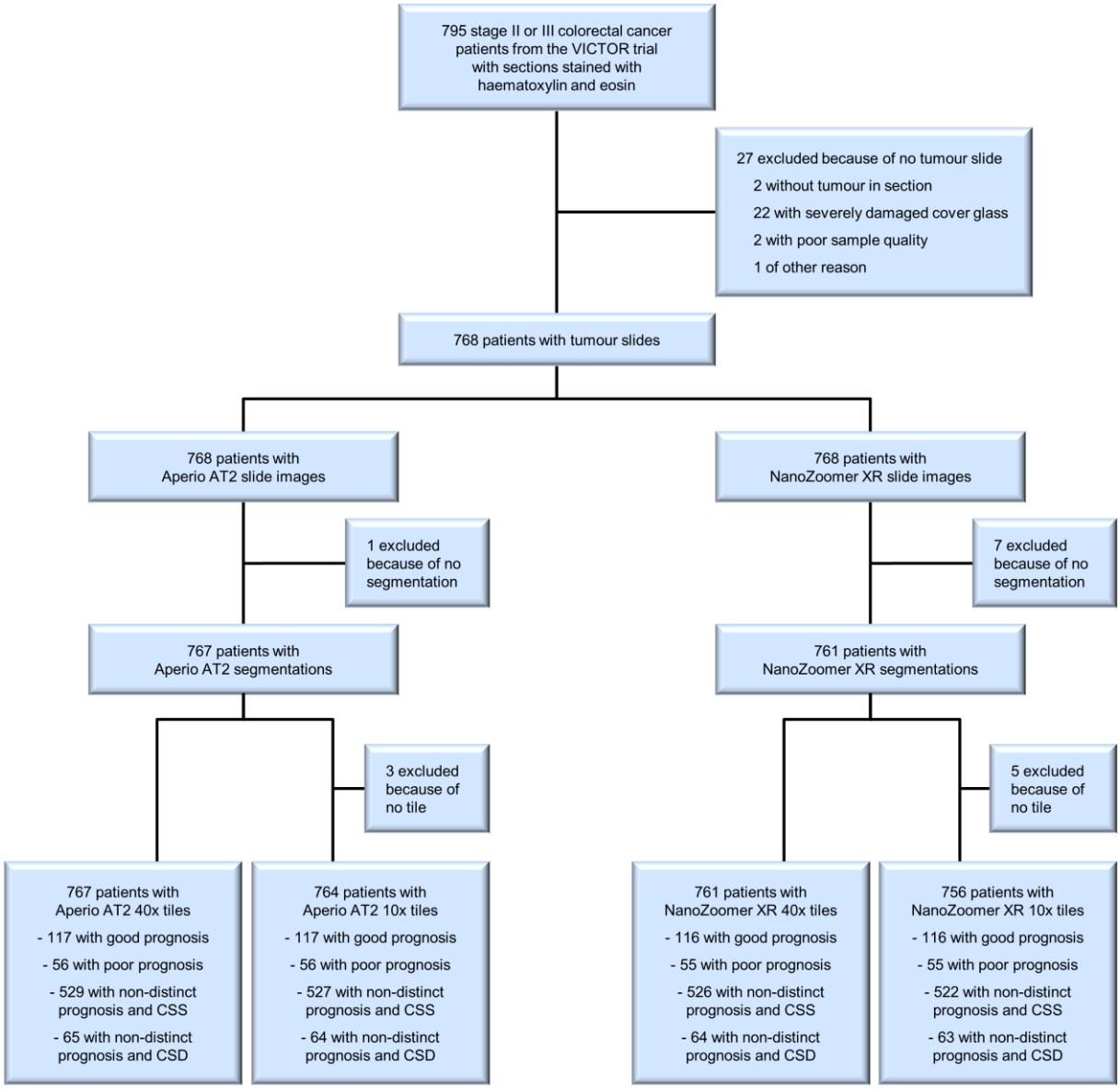
same manner as for the Ahus cohort, resulting in 969 patients with Aperio AT2 segmentations and 967 patients with NanoZoomer XR segmentations (Protocol Fig. 3). These patients constituted the Gloucester cohort, but one of them was excluded from the Aperio AT2 10x tile set because of no tile within the tumour segmentation (Protocol Fig. 3).



Protocol Fig. 3 | A diagram specifying inclusions and exclusions of patients, slides and slide images from the Gloucester cohort, and the prognosis of the included patients. CSS, cancer-specific survival; CSD, cancer-specific death.

2.4 VICTOR cohort

The VICTOR trial randomised stage II and III colorectal cancer patients to receive rofecoxib or placebo after primary treatment in order to examine cardiovascular adverse events[5,6]. An H&E-stained 3 µm section from a FFPE tissue block was retrieved for 795 of the patients recruited between 2002 and 2004, some of which were sectioned at ICGI and some of which were sectioned elsewhere. The sections were processed in the same manner as for the Ahus cohort. The VICTOR cohort consisted of 767 patients with Aperio AT2 40x tiles, 764 patients with Aperio AT2 10x tiles, 761 patients with NanoZoomer XR 40x tiles and 756 patients with NanoZoomer XR 10x tiles (Protocol Fig. 4).



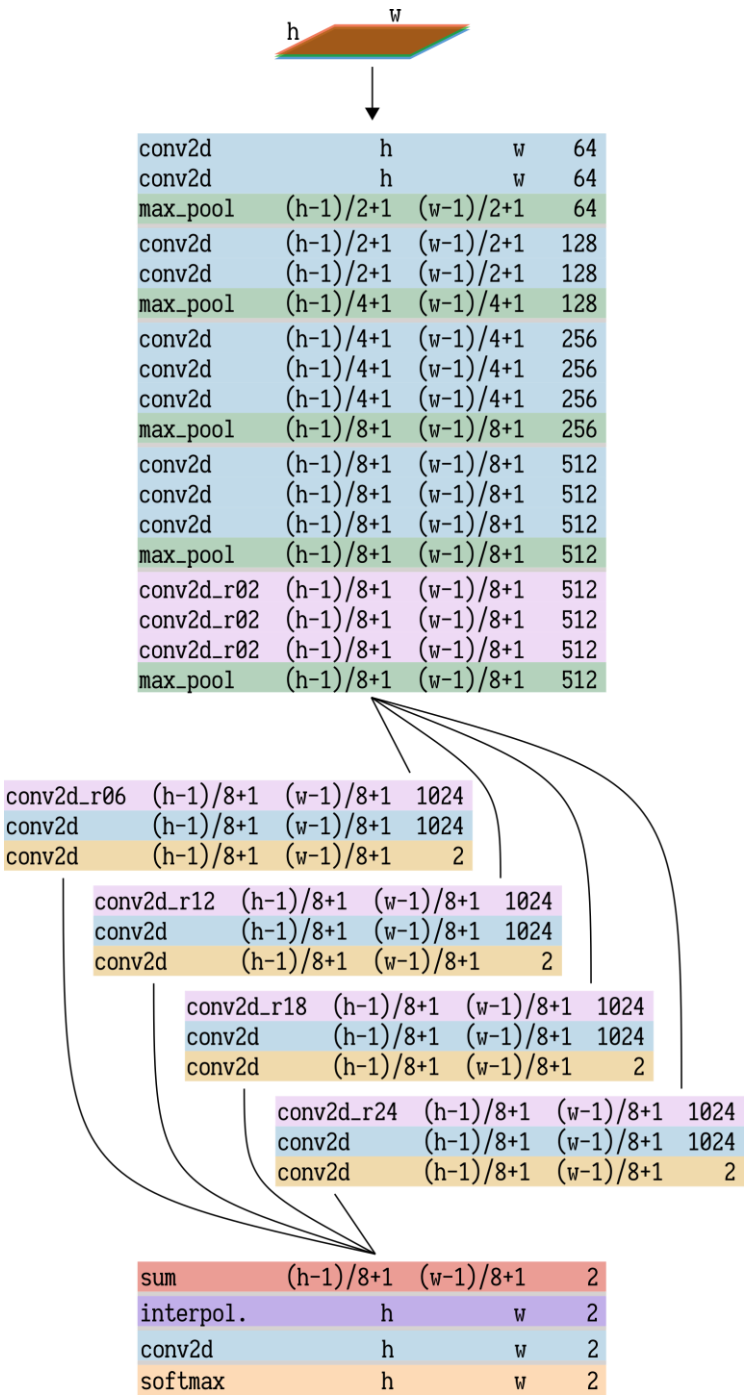
Protocol Fig. 4 | A diagram specifying inclusions and exclusions of patients, slides and slide images from the VICTOR cohort, and the prognosis of the included patients. CSS, cancer-specific survival; CSD, cancer-specific death.

3 Segmentation

The segmentation method consists of a method to produce probability maps from input images, and a different method to create an image partitioned into foreground and background regions based on the input image and the corresponding probability map. The probability maps are generated by a segmentation network based on the DeepLab network[7] (Protocol Fig. 5). The final segmentation is achieved using dense conditional random fields[8].

The method was initially trained on 1077 images with corresponding annotations from the Aker cohort (670 images) and VICTOR cohort (407 images). The images were obtained from slides scanned with a NanoZoomer (Hamamatsu Photonics, Japan) scanner, and the annotations was hand-drawn by a pathologist. This trained method was then applied on images from the Aker cohort, the Ahus cohort, and the Gloucester cohort. The resulting segmentations was verified by a pathologist, which corrected the ones that was unsatisfactory. This set of images with corresponding (possibly corrected) masks constitutes the development dataset of the image segmentation method.

From the development dataset of 1717 patients, 25% (429 patients) was drawn uniformly at random to form a tuning set, and the remaining 1288 patients comprised the training set. In the training set there was 358 patients with a cancer-specific event, and 930 images patients without. In the tuning set there was 128 images patients with a cancer-specific event, and 301 patients without. Slides from patients in the segmentation development set was scanned with both an Aperio AT2 scanner and a NanoZoomer XR scanner. The development set in the segmentation task is therefore comprised of 3430 scans (4 scans from the NanoZoomer XR scanner was missing), with 2573 in the training set, and 857 in the tuning set.



Protocol Fig. 5 | Illustration of the segmentation network architecture. Each layer is represented by name, output height, output width, and number of output channels. Progression is downwards from the input image at the top to the prediction output at the bottom.

Each scan is digitally resized to a size corresponding to a 2.5x resolution (see section 4), and stored as a PNG image. Each image is then resized to fit within a frame of 1600x1600 pixels with a Catmull-Rum cubic filter. This is done by resizing the image while preserving the aspect ratio until its largest dimension (height or width) is 1600 pixels. A new image is then formed by padding the resized image along its shortest dimension on each

side until it also is 1600 pixels. The centre of the resized image aligns with the centre of the padded image, and the padded image is used further.

The segmentation network was trained with 100,000 update steps (training iterations), and each update step uses 16 images (this collection is called a mini-batch) distributed on 4 GPUs. Every image in the development dataset is used once before one is used twice, which means that each image is seen on about 622 times during training (one progression through a dataset is termed an epoch). At each epoch, the same image is used once, but with slight variations each time. First, a section of 641x641 pixels is cropped at a random location within the image. Then, a set of orientation distortions are applied in the following order

1. With a probability of 50%, flip the image horizontally (mirror along its horizontal axis).
2. With a probability of 50%, flip the image vertically (mirrored along its vertical axis).
3. With a probability of 50%, rotate the image once with one of the following degrees: 0, 90, 180, 270.

Finally, the image is centred around its mean and standard deviation (see https://www.tensorflow.org/versions/r1.10/api_docs/python/tf/image/per_image_standardization). The resulting image is fed into the segmentation network as an RGB image.

The trainable parameters are initialized using a Xavier weight initialization scheme, and updated using a standard stochastic gradient descent optimization method[9]. The step length in the optimization is initialized to 0.05, and decreased by a factor of 0.1 at iteration 96488 (about 600 training epochs).

Applying the trained network on an image yields a probability map with the same spatial shape as the image. This probability map is a one-channel grayscale image with intensity values in 0, 1, …, 255. The method assigns high values to regions it finds probable depicting cancerous tissue.

For each image, we create additional versions of the image by rotating and flipping the original image, before we apply the trained network on all the different versions. There are 8 versions, and they are obtained from the original image by the following operations

1. Do nothing (this is the original image)
2. Flip the image around its horizontal axis
3. Flip the image around its vertical axis
4. Rotate the image 90 degrees clockwise
5. Rotate the image 180 degrees clockwise
6. Rotate the image 270 degrees clockwise
7. Rotate the image 90 degrees clockwise and flip the result around its horizontal axis
8. Rotate the image 270 degrees clockwise and flip the result around its horizontal axis

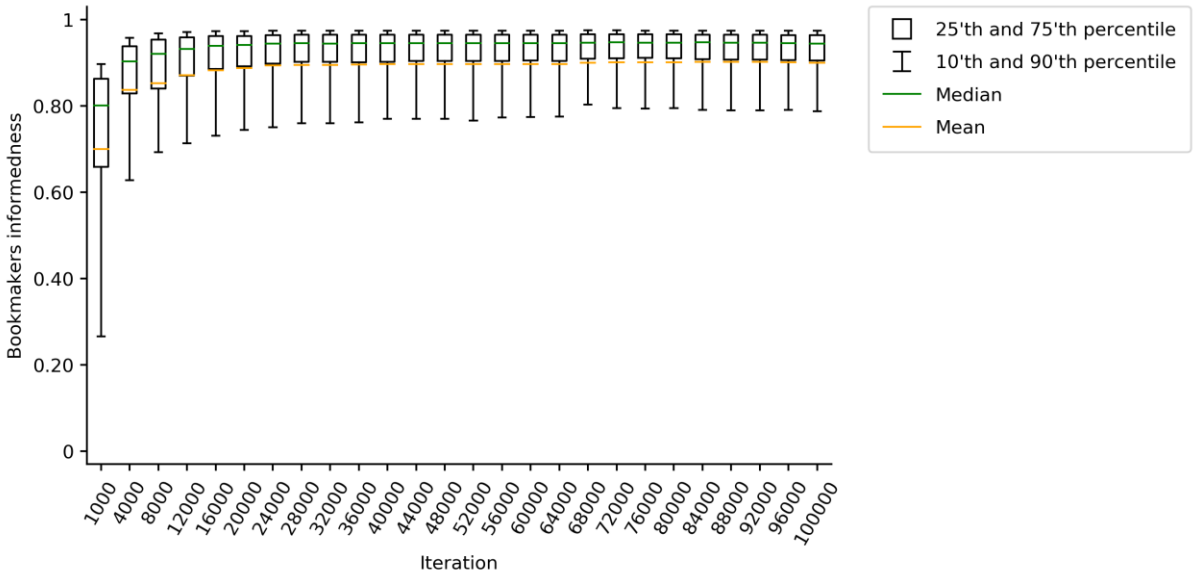
The resulting probability maps are then restored to their original orientation, and an average image of all the different versions is computed and used further in the process.

At inference, the trained network is applied on one image at the time (i.e. with a batch-size of one), and contrary to the training phase, neither cropping nor orientation distortion is applied. However, it is important that every image is centred around its mean and standard deviation as was done in training. The network was implemented and run in Python 3.5 (https://www.python.org) using TensorFlow 1.10 (https://www.tensorflow.org).

Segmentation of the probability maps was performed using the Python library pydensecrf v1.0rc3 (https://github.com/lucasb-eyer/pydensecrf). The model used a unary potential (the probability map), a gaussian pairwise potential (addPairwiseGaussian(sxy=1, compat=1)), and a bilateral pairwise potential (addPairwiseBilateral(sxy=30, srgb=3, compat=100)). The result image with float values in (0, 1) is thresholded at 0.5 to produce a binary mask, where pixels with value less than 0.5 is labelled as background, and the rest as foreground.

The resulting segmentation is smoothed with a 5x5 mean filter, before foreground regions connected with a eight-neighbourhood with fewer than 20,000 pixels are removed. Background regions fully contained within foreground regions are marked as foreground.

The method was applied on the tuning set every 4,000 iterations, and the predicted segmentations was evaluated against the reference segmentations. The model that achieved the highest mean bookmakers informedness score was then selected as the model to be used in the rest of the experiment. The model at iteration 88,000 achieved the highest score of 0.902 (Protocol Fig. 6).



Protocol Fig. 6 | Performance of the segmentation method on the tuning set. The method is evaluated at multiple training iterations evenly spaced across the training progression.

4 Tiling

The region identified as tumour by the segmentation method is not directly suitable as input to a convolutional neural network (CNN) because of limited GPU memory in commonly available hardware. We therefore made multiple non-overlapping regions of a fixed size, called *tiles*, from within the region segmented as tumour in each slide image. Since the physical area represented by a pixel depends on the scanner[*], tiles representing the

[*] The physical area represented by a pixel can also depend on the applied scan settings, but this is not an issue here as we used the same settings for each of the scanners when scanning the slides in the training and validation cohorts.

same physical area were created by including slightly different number of pixels in tiles from Aperio AT2 and NanoZoomer XR slide images. At maximum resolution, termed *40x*, pixels in the Aperio AT2 slide images had a physical size of 0.253 µm/pixel both vertically and horizontally, while pixels in the NanoZoomer XR slide images had 0.227 µm/pixel both vertically and horizontally. To make 40x tiles, tiles with 486x486 pixels were extracted from within the tumour segmentation of Aperio AT2 slide images, while 542x542 pixels were used for NanoZoomer XR slide images. Similarly, a tile size of 1942x1942 pixels were used for Aperio AT2 slide images and 2166x2166 pixels for NanoZoomer XR slide images to make 10x tiles. Each of these raw tiles was then resampled to 512x512 pixels, making the physical area of each pixel similar for both scanners; 0.240x0.240 µm for 40x tiles and 0.960x0.960 µm for 10x tiles.

Technically, the tiling was performed by defining a grid of candidate tiles from the top left corner of the slide image, including regions outside the tumour segmentation. Candidate tiles for which the four corners and their midpoints along the edges were within the boundaries of the segmentation were included. Tiles were extracted with OpenSlide from level 0, converted to numpy arrays, resized with OpenCV using the resize() function (https://docs.opencv.org/3.4.0/da/d54/group__imgproc__transform.html) with interpolation set to cv2.INTER_CUBIC for up-sampling and cv2.INTER_AREA for down-sampling and saved in a lossless format (as PNG files).

5 Patient survival prediction methods

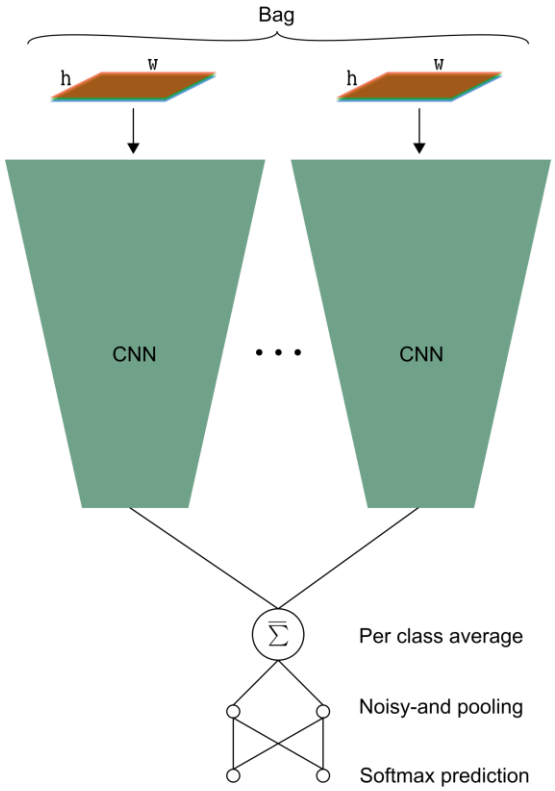
Two neural networks were trained using all patients with distinct prognosis in the training cohorts. Each network was trained five times with 40x tiles and another five times with 10x tiles; the resampled tiles with 512x512 pixels were used in both cases. The applied ground truth (i.e. true outcome) in these supervised classification methods was the patient's distinct prognosis, either good or poor prognosis (as defined in section 2).

5.1 DoMore v1 network

One network, called DoMore v1 network, is a multiple instance classification method comprised of a representation network, a multiple instance pooling function, and a classification network (Protocol Fig. 7).

Rather than a single tile, the DoMore v1 network classifies a collection of tiles called a bag, where all tiles within a bag originates from the same scan image. Each tile in a bag is applied with the representation network to produce a feature representation of the tile (note that within one update step, all tiles uses the same representation network with the same parameter values). All tile representations are aggregated and a single value for each class is produced by the pooling function. A final classification network is then applied, and predictions are produced. These predictions are compared with the ground truth corresponding to the image the bag originates from using a loss function. This loss function is optimised using a gradient-based optimisation routine, and at each training iteration, the trainable parameters of the network are updated according to this optimisation method. Only a randomly selected subset of the tiles in the bag is used to update the network. This asymmetric forward and backward propagation reduces the memory footprint of the network, and allows larger bags of tiles during training.

The representation network is based on the MobileNet v2 network, and its details is illustrated in Protocol Fig. 7[10]. The first convolution layer in the representation network uses a 3x3 convolution kernel with a stride of 2. The activation function is a ReLU activation function[11]. Inside each inverted bottleneck module, the first convolution layer uses a 1x1 convolution kernel with a stride of 1 and a ReLU6 activation function[12]. The depth-wise separable convolution layer uses a 3x3 convolution kernel. Whenever the spatial size halves in height and width, the stride is 2, otherwise it is 1. The activation function is the ReLU6 function. The last convolution layer uses a 1x1 convolution kernel with a stride of 1, and an identity activation function. When the number of input channels to the inverted bottleneck module is equal to the number of output channels in the same module, the input to the first convolution layer within the module is added to the result of the last convolution layer within the module. The convolution layer after the inverted bottleneck modules uses a 1x1 convolution with stride 1 and the ReLU activation function. All convolution and separable convolution layers described above employ batch normalization on the result of the convolution, before the activation function is applied[13]. All kernel weights are initialised with Xavier initialization, and no bias parameters are used. The final convolution layer uses a 1x1 convolution kernel with stride 1. No batch normalization is used in this layer, and the activation is the identity function. The rest of the network consists of a noisy-and pooling function followed by a softmax classification, following the design of Kraus and colleagues[14]. There is one cross-entropy loss function associated with the output of the pooling function, and one cross-entropy loss function associated with the classification output.



CNN

conv2d	h/ 2	w/ 2	32	
conv2d	h/ 2	w/ 2	32	inverted bottleneck
separable_conv2d	h/ 2	w/ 2	32	
conv2d	h/ 2	w/ 2	16	
conv2d	h/ 2	w/ 2	96	inverted bottleneck
separable_conv2d	h/ 4	w/ 4	96	
conv2d	h/ 4	w/ 4	24	
conv2d	h/ 4	w/ 4	144	inverted bottleneck
separable_conv2d	h/ 4	w/ 4	144	
conv2d	h/ 4	w/ 4	24	
conv2d	h/ 4	w/ 4	144	inverted bottleneck
separable_conv2d	h/ 8	w/ 8	144	
conv2d	h/ 8	w/ 8	32	
conv2d	h/ 8	w/ 8	192	inverted bottleneck
separable_conv2d	h/ 8	w/ 8	192	
conv2d	h/ 8	w/ 8	32	
conv2d	h/ 8	w/ 8	192	inverted bottleneck
separable_conv2d	h/ 8	w/ 8	192	
conv2d	h/ 8	w/ 8	32	
conv2d	h/ 8	w/ 8	192	inverted bottleneck
separable_conv2d	h/16	w/16	192	
conv2d	h/16	w/16	64	
conv2d	h/16	w/16	192	inverted bottleneck
separable_conv2d	h/16	w/16	192	
conv2d	h/16	w/16	64	
conv2d	h/16	w/16	384	inverted bottleneck
separable_conv2d	h/16	w/16	384	
conv2d	h/16	w/16	64	
conv2d	h/16	w/16	384	inverted bottleneck
separable_conv2d	h/16	w/16	384	
conv2d	h/16	w/16	64	
conv2d	h/16	w/16	384	inverted bottleneck
separable_conv2d	h/16	w/16	384	
conv2d	h/16	w/16	96	
conv2d	h/16	w/16	576	inverted bottleneck
separable_conv2d	h/16	w/16	576	
conv2d	h/16	w/16	96	
conv2d	h/16	w/16	576	inverted bottleneck
separable_conv2d	h/16	w/16	576	
conv2d	h/16	w/16	96	
conv2d	h/16	w/16	576	inverted bottleneck
separable_conv2d	h/32	w/32	576	
conv2d	h/32	w/32	160	
conv2d	h/32	w/32	960	inverted bottleneck
separable_conv2d	h/32	w/32	960	
conv2d	h/32	w/32	160	
conv2d	h/32	w/32	960	inverted bottleneck
separable_conv2d	h/32	w/32	960	
conv2d	h/32	w/32	160	
conv2d	h/32	w/32	960	inverted bottleneck
separable_conv2d	h/32	w/32	960	
conv2d	h/32	w/32	320	
conv2d	h/32	w/32	1280	
avg_pool2d	1	1	1280	
conv2d	1	1	2	

Protocol Fig. 7 | Illustration of the DoMore v1 network architecture. The left side gives an overview over progression from an input bag of tiles to a bag prediction. The right side shows the architecture of the representation network, where each layer is represented by name, output height, output width, and number of output channels.

The network was trained with a batch size of 32 bags, distributed across 8 GPUs with 4 bags on each GPU. Each bag consisted of 64 tiles with size 512x512x3 pixels and with values in 0, 1, …, 255. The number of tiles contributing to the gradient computation was 8. For updating the network parameters, an initial step size of 0.001 was used with the Adam optimisation method[15]. When training on 10x tiles, the learning rate was initially set to 0.001 and then successively reduced by a factor of 0.1 at iteration 6,000 and again at iteration 12,000 before training ceased after iteration 15,000. Twice the number of iterations were utilised to train on 40x tiles, i.e. the learning rate started at 0.001 and was successively reduced by a factor of 0.1 at iteration 12,000 and again at iteration 24,000 before training ceased after iteration 30,000.

At each step, before entering the network, each tile is distorted and normalised. First, it is randomly cropped to a size of 448x448, before the orientation of the tile is distorted. The tile is randomly flipped from left to right (around its central vertical axis), then randomly flipped from top to bottom (around its central horizontal axis), and finally randomly rotated by either 0°, 90°, 180° or 270°. Then its values are scaled to (0, 1) by casting it to a 32-bit floating point number before dividing the entire tile by 255.0. The tile is then converted from the RGB colour space to the HSV colour space before each channel is scaled with a value uniformly distributed between 1/1.1 and 1.1. The tile is then converted back to RGB. Finally, the tile is normalised to have zero mean and unit norm (see rgb_to_hsv, hsv_to_rgb, per_image_standardization at https://www.tensorflow.org/versions/r1.10/api_docs/python/tf/image for more information). At inference, no cropping is applied, so the entire tile of size 512x512x3 pixels is evaluated by the network. Also no orientation or colour distortions are applied. Before entering the network, each tile is normalised to have zero mean and unit norm as in training. The network was implemented and run in Python 3.5 (https://www.python.org) using TensorFlow 1.10 (https://www.tensorflow.org). To account for class imbalance in the training set, the minority class within a cohort-scanner combination was oversampled such that there was an equal amount of images labelled with good prognosis and poor prognosis in every cohort-scanner combination. Within each cohort-scanner combination, images were sampled uniformly at random without replacement.

5.2 Inception v3 network

The other network, an Inception v3 network[16], was trained with Keras (2.1.6) using the Tensorflow Docker image (tensorflow/tensorflow:1.9.0-gpu-py3). The input image size was 512x512 and the output was two classes with the first class being the probability of good prognosis and the second class the probability of poor prognosis. A binary cross entropy loss function was used, and it was optimised with keras.optimizers.Adam using default arguments, except for initial learning rate which was set to 0.0001. To account for class imbalance between tiles from good and poor prognosis, tiles from the minority class were oversampled per cohort prior to training and the file paths were saved as a list. Consequently, each cohort contained the same number of included tiles with good and poor prognosis, at the expense of potentially including some tiles twice. The list of tiles was loaded prior to training and randomly shuffled before a modified version of keras.preprocessing.image.ImageDataGenerator was utilised to load batches of images using 16 worker threads. The ImageDataGenerator was modified to perform colour distortion by

1. converting the tile to HSV colour space,
2. augmenting the hue by adding a random uniformly sampled value between ±0.05,
3. scaling the saturation by a random uniformly sampled value between 1/1.1 and 1.1,
4. shifting the saturation by a random uniformly sampled value between ±0.1,
5. scaling the value by a random uniformly sampled value between 1/1.1 and 1.1,
6. shifting the value by a random uniformly sampled value between ±0.1, and
7. converting the tile back to the RGB colour space.

The tile was then standardised by subtracting the mean colour values and dividing by the standard deviation of all tiles used for training, i.e. all tiles of patients with distinct prognosis in the training cohorts. For each training iteration, a batch size of 16 tiles was used due to GPU memory constraints. When training on 10x tiles, the learning rate was initially set to 0.0001 and then successively halved for each 25,000nd iteration, starting at iteration 25,000, before training ceased after iteration 150,000. Twice the number of iterations were utilised to train on 40x tiles, i.e. the learning rate started at 0.0001 and was successively halved for each 50,000nd iteration, starting at iteration 50,000, before training ceased after iteration 300,000. The network output was the predicted

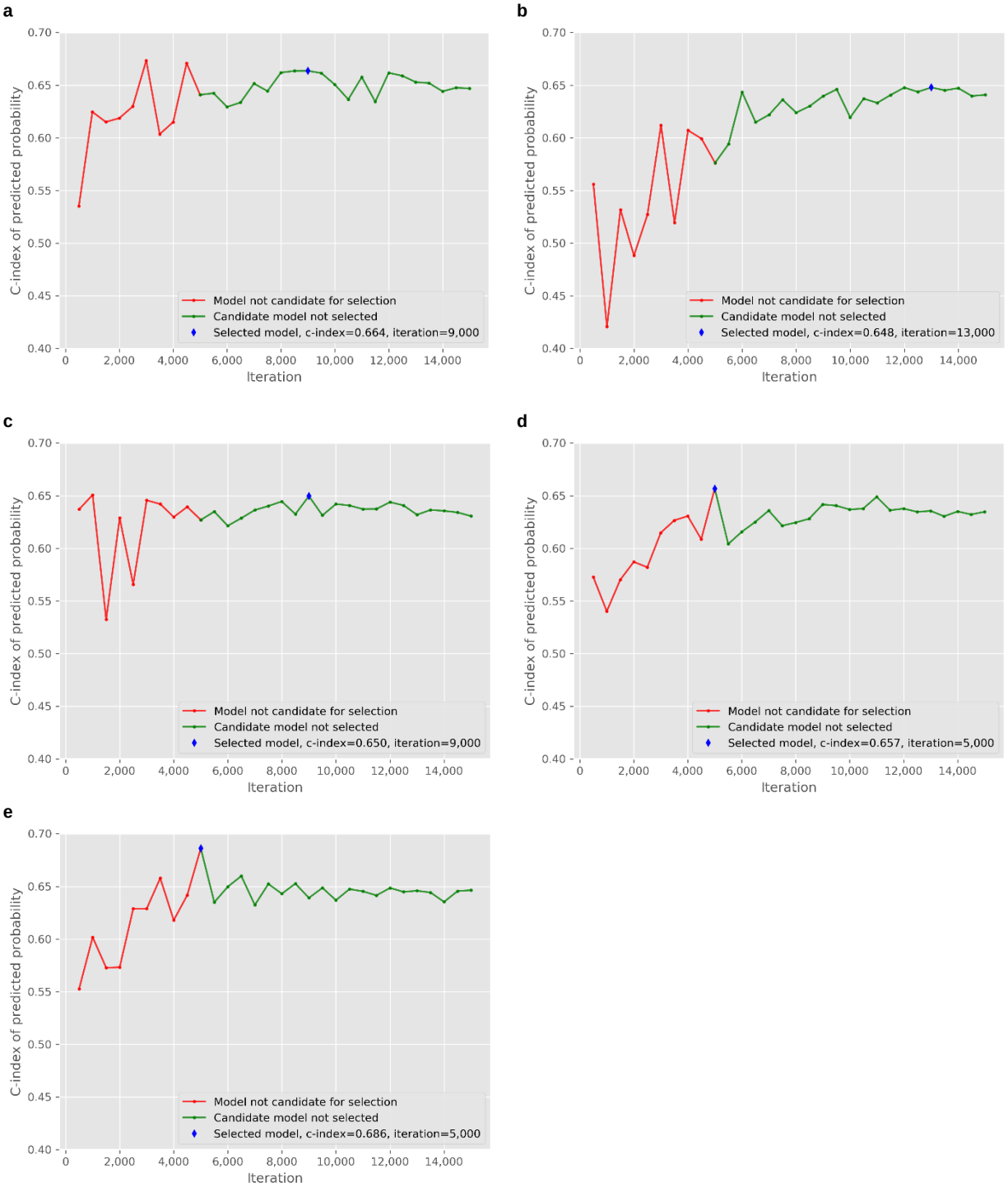
probability of poor prognosis for a tile. The predicted probability of poor prognosis for a patient was computed by averaging the predicted probabilities of all tiles for that patient.

6 Individual models

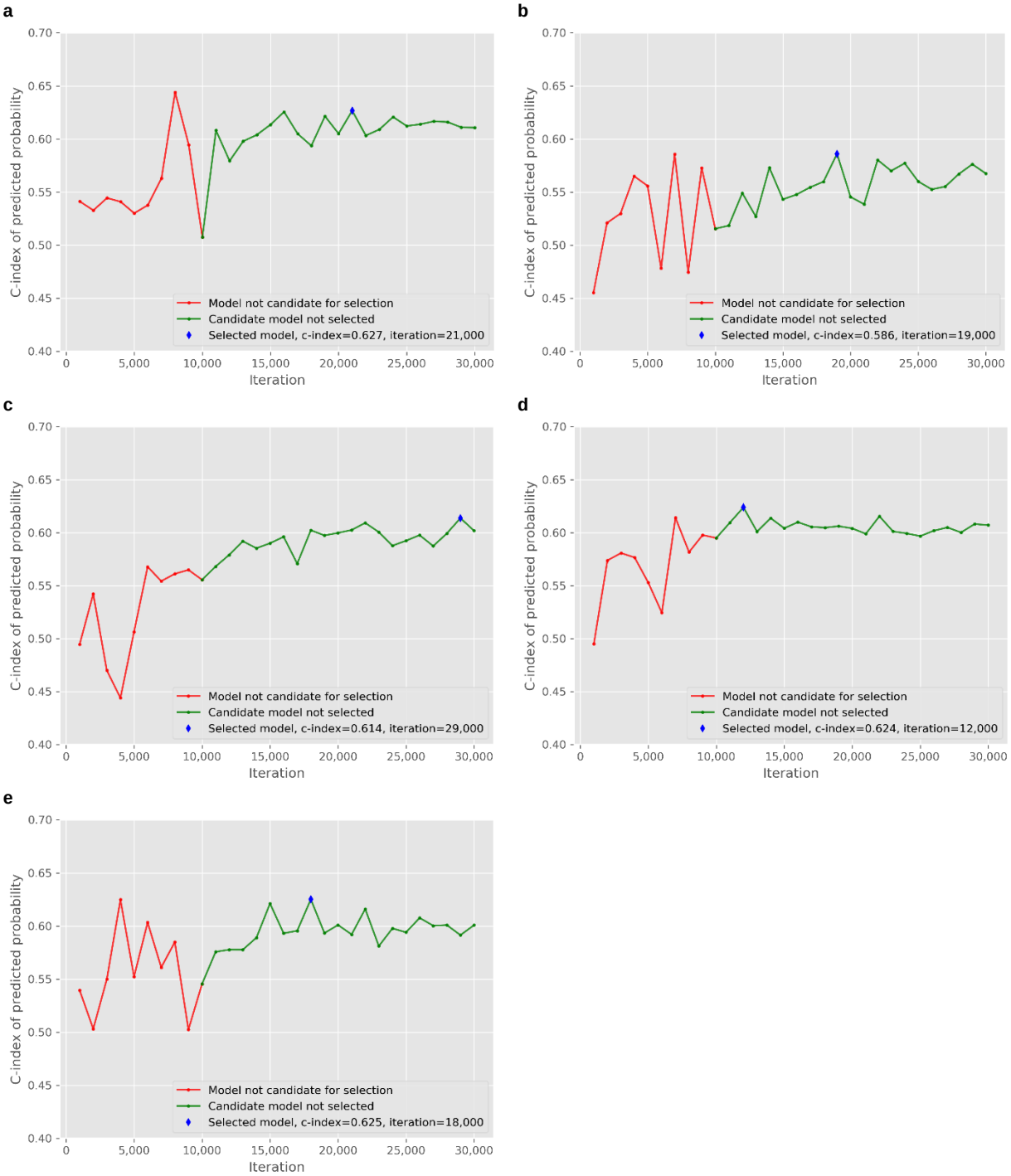
Training each of the two networks five times for each of the two resolutions resulted in 20 training runs. For each of these 20 training runs, 21 models were evaluated on all patients with non-distinct prognosis in the training cohorts. The 21 models evaluated for each training run was uniformly distributed from 1/3 of the iterations to the training ceased (both ends inclusive). Each 10x model of the DoMore v1 network was evaluated at iteration 5,000, 5,500, and so on up to iteration 15,000. Each 40x model of the DoMore v1 network was evaluated at iteration 10,000, 11,000, and so on up to iteration 30,000. Each 10x model of the Inception v3 network was evaluated at iteration 50,000, 55,000, and so on up to iteration 150,000. Each 40x model of the Inception v3 network was evaluated at iteration 100,000, 110,000, and so on up to iteration 300,000.

To reduce evaluation time for the 40x models, a random sample of 2,000 40x tiles were selected for each slide with more than 2,000 40x tiles. The same tiles were evaluated for all models. To reduce further the evaluation time for the 40x models of the DoMore v1 network, patients with more than 50 tiles were evaluated using 50 tiles at a time, resulting in that tiles ordered after the last multiple of 50 were ignored in these evaluations, i.e. at most 49 tiles were ignored for each patient. Note that these speed-ups were only applied during model selection; for all applications of the selected models, including the external evaluation described in this protocol, all tiles will be evaluated.

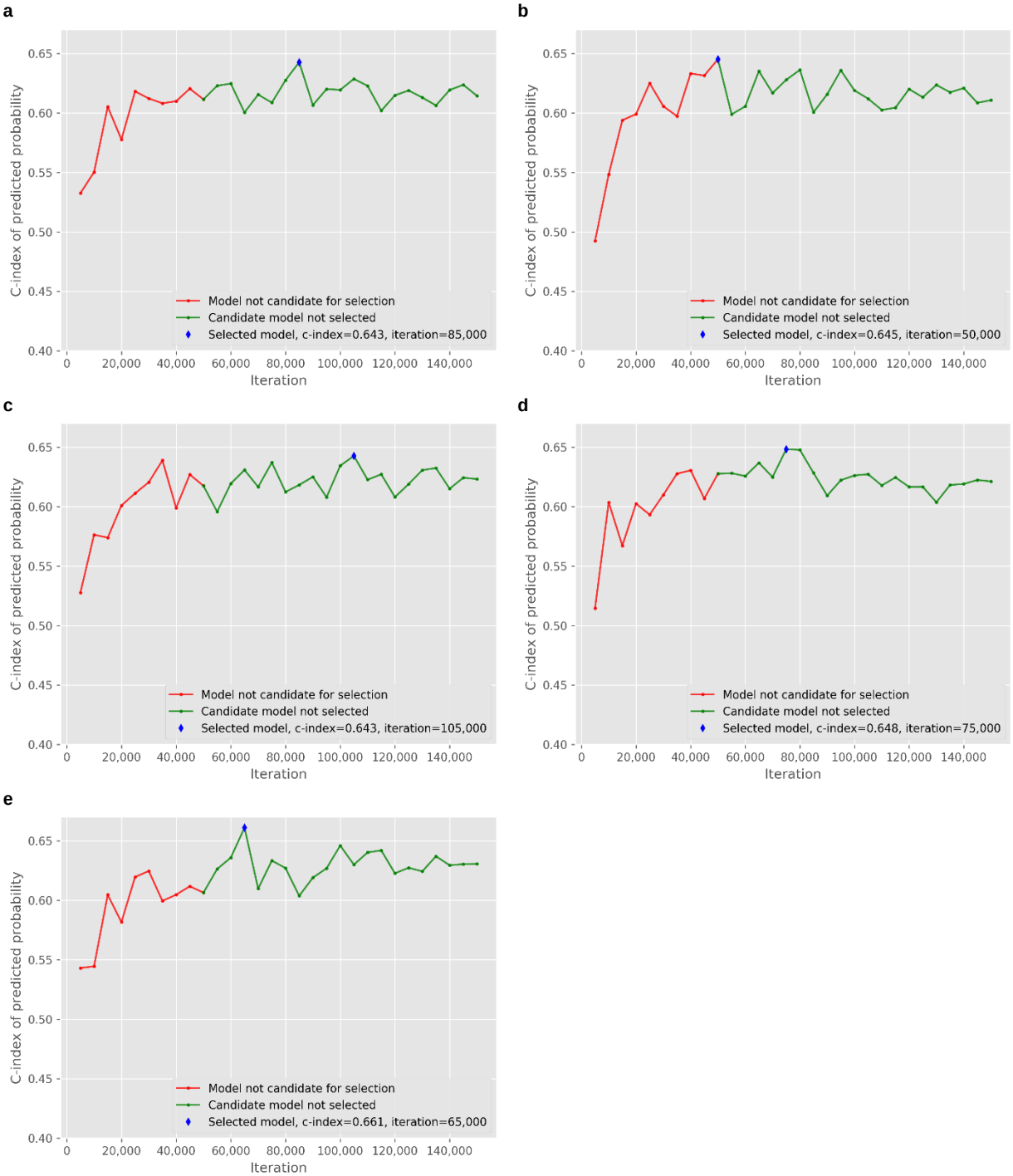
The model that maximised Harrell's concordance index[17] (c-index) was selected for each training run. The c-index compared the observed time to cancer-specific death or censoring to a model's predicted probability of poor prognosis for patients with non-distinct prognosis in the training cohorts. The model with largest c-index thus appeared to provide most prognostic information in its predicted probabilities when evaluated on non-distinct prognosis patients in the training cohorts. Protocol Figs. 8-11 show the c-index of all candidate models and indicate the selected model for each of the 21 training runs.



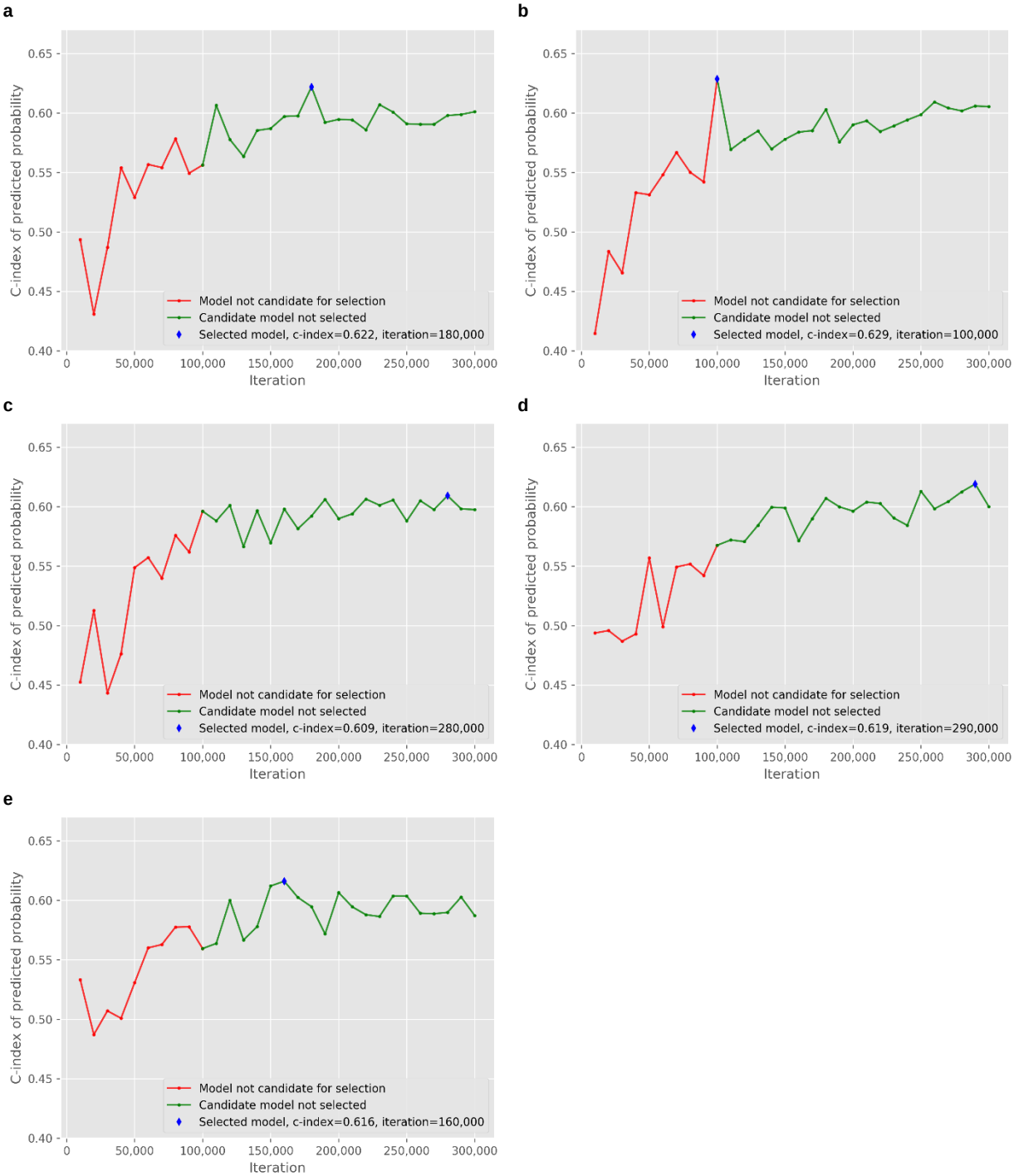
Protocol Fig. 8 | c-index of the 21 candidate 10x models of the DoMore v1 network for patients with non-distinct prognosis in the training cohorts. The blue point indicates the selected model, green points indicate models not selected. The c-index of nine models from the first third of the training run is shown as red points for comparison. Subplot **a** to **e** show training run 1 to 5.



Protocol Fig. 9 | c-index of the 21 candidate 40x models of the DoMore v1 network for patients with non-distinct prognosis in the training cohorts. The blue point indicates the selected model, green points indicate models not selected. The c-index of nine models from the first third of the training run is shown as red points for comparison. Subplot **a** to **e** show training run 1 to 5.



Protocol Fig. 10 | c-index of the 21 candidate 10x models of the Inception v3 network for patients with non-distinct prognosis in the training cohorts. The blue point indicates the selected model, green points indicate models not selected. The c-index of nine models from the first third of the training run is shown as red points for comparison. Subplot **a** to **e** show training run 1 to 5.

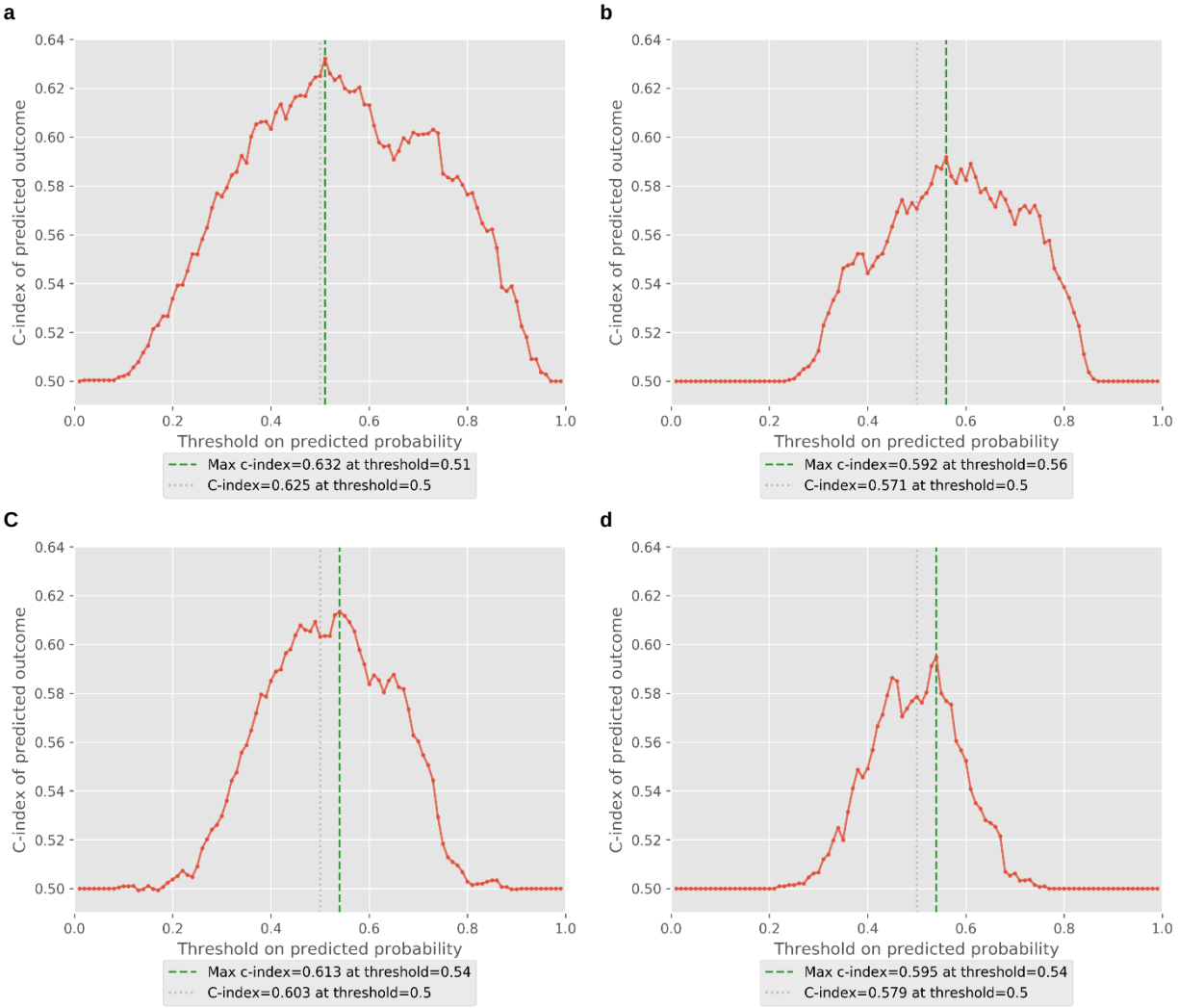


Protocol Fig. 11 | c-index of the 21 candidate 40x models of the Inception v3 network for patients with non-distinct prognosis in the training cohorts. The blue point indicates the selected model, green points indicate models not selected. The c-index of nine models from the first third of the training run is shown as red points for comparison. Subplot **a** to **e** show training run 1 to 5.

7 Ensemble models

An ensemble model was created for each network and resolution by averaging the five selected models' predicted probability of poor prognosis for a patient, resulting in four ensemble models; a 10x and a 40x ensemble model of the DoMore v1 network and similarly for the Inception v3 network. To determine a suitable

threshold for dichotomising each ensemble model's predicted probability of poor prognosis, we computed the c-index of the dichotomised ensemble model prediction for thresholds at 0.01, 0.02, and so on up to and including 0.99 for patients with non-distinct prognosis in the training cohorts. The threshold obtaining the maximum c-index was selected for each ensemble model (Protocol Fig. 12). For the 10x ensemble model of the DoMore v1 network, the predicted outcome was poor prognosis if the ensemble model's predicted probability of poor prognosis was greater than 0.51, otherwise, the predicted probability was less than or equal to 0.51 and the predicted outcome was good prognosis. This dichotomous marker was termed the 10x ensemble marker of the DoMore v1 network. Similarly, the 40x ensemble marker of the DoMore v1 network was defined using a threshold of 0.56, and both the 10x and the 40x ensemble marker of the Inception v3 network was defined using a threshold of 0.54.



Protocol Fig. 12 | c-index of an ensemble model's predicted probability of poor prognosis thresholded at 0.01, 0.02, and so on up to and including 0.99 for patients with non-distinct prognosis in the training cohorts. **a**,The 10x ensemble model of the DoMore v1 network. **b**,The 40x ensemble model of the DoMore v1 network. **c**,The 10x ensemble model of the Inception v3 network. **d**,The 40x ensemble model of the Inception v3 network.

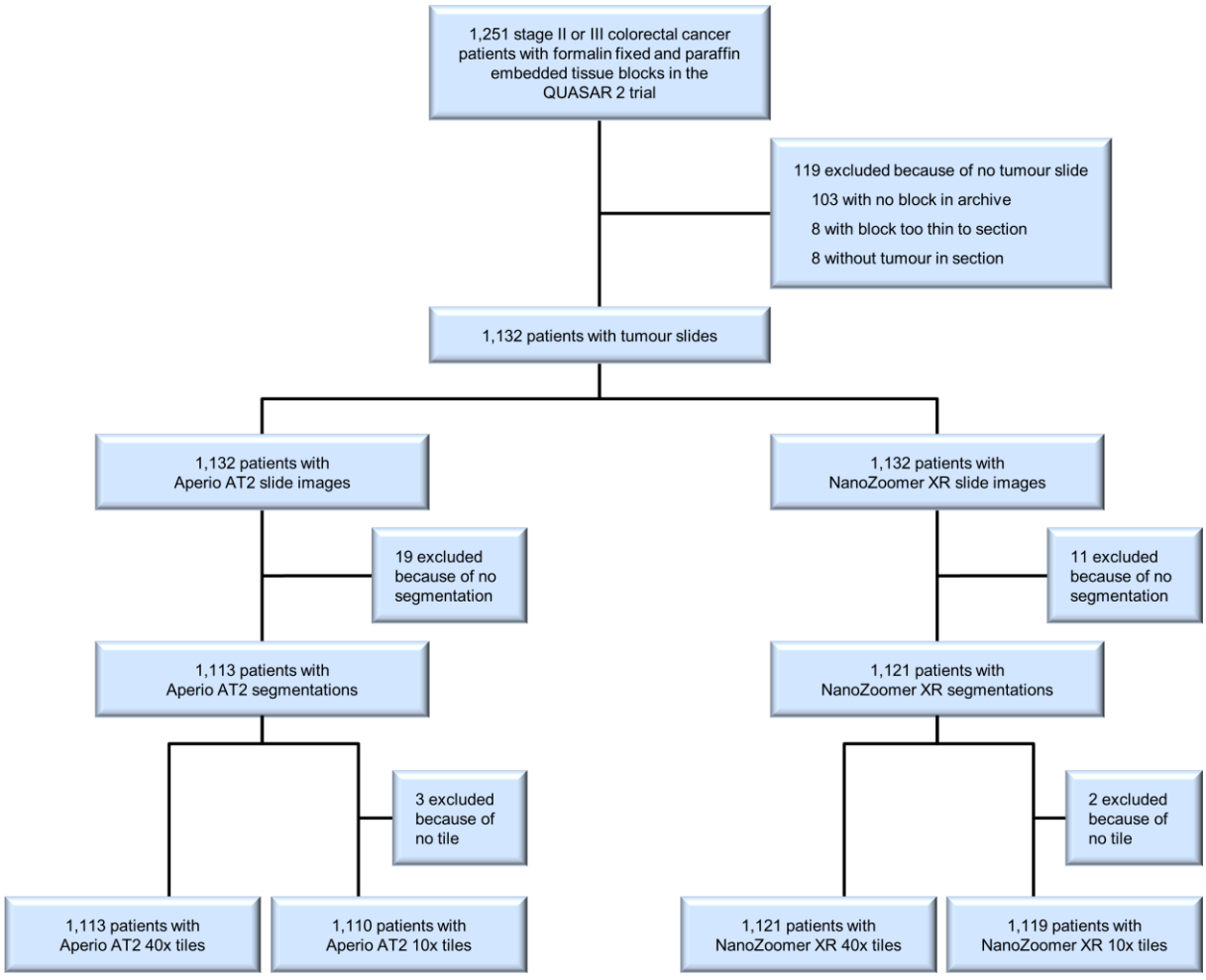
A combined 10x and 40x ensemble model was created for each network by defining patients were the 10x and 40x ensemble markers predict the same outcome as either predicted good prognosis (if both ensemble markers predict good prognosis) or predicted poor prognosis (if both ensemble markers predict poor prognosis), and patients where the 10x and 40x ensemble markers predict different outcome as predicted uncertain prognosis. If one of the ensemble markers could not be assayed for a patient, e.g. because there was no 10x tiles, then the combined 10x and 40x ensemble marker was not defined either, thus such patients was excluded from analyses of the combined model. This resulted in two combined 10x and 40x ensemble markers, one for the DoMore v1 network and one for the Inception v3 network. These 3-grouped variables will be referred to as the DoMore v1 marker and the Inception v3 marker.

8 QUASAR 2 cohort

The open-label, randomised, controlled QUASAR 2 trial (ISRCTN registry number ISRCTN45133151) enrolled 1952 patients with histologically proven stage III or high-risk stage II colorectal cancer between April 2005 and October 2010 from 170 hospitals in seven countries (Australia, Austria, Czech Republic, New Zealand, Serbia, Slovenia and the UK), of whom 1,941 had assessable data[18]. The trial was designed to investigate whether bevacizumab improved disease-free survival after potentially curative surgery of primary tumour. All patients received adjuvant chemotherapy in the form of capecitabine, but none received neoadjuvant treatment. No significant difference was observed between the treatment arms and the investigators concluded that the addition of bevacizumab to capecitabine should not be used in this adjuvant setting[18].

Through encouraging, but not requiring blood samples and tumour samples from primary resections, FFPE tissue blocks were collected from 1,251 of the QUASAR 2 trial patients with stage II or III colorectal cancer. These patients were representative for the whole trial population in terms of clinical and pathological characteristics[18]. Pathologists at the participating hospitals in the trial performed the pathological evaluations. All patients provided written informed consent for treatment and the use of tissue samples. The West Midlands Research Ethics Committee (no. 04/MRE/11/18) and the Regional Committees for Medical and Health Research Ethics (REK) in Norway (no. 2015/1607) approved the study.

Tissue blocks from 1,140 patients were received, sectioned and prepared as 3 µm H&E-stained tissue slides by laboratory personnel at ICGI (Protocol Fig. 13). A local pathologist blinded to clinical outcome ascertained the presence of tumour in each tissue section. Digital images of the 1,132 sections with tumour were acquired using the same two scanners as in the training cohorts, i.e. an Aperio AT2 and a NanoZoomer XR. The previously developed segmentation model was blindly applied to automatically identify regions with tumour, giving 1,113 patients with Aperio AT2 segmentations and 1,121 patients with NanoZoomer XR segmentations (Protocol Fig. 13). The slide images where tiled as in the training cohorts, but no 10x tile could fit inside the automatic tumour segmentation for 3 Aperio AT2 segmentations and 2 NanoZoomer XR segmentations (Protocol Fig. 13). The QUASAR 2 cohort was defined as the 40x tiles from the Aperio AT2 slide images (available for 1,113 patients), the 10x from the Aperio AT2 slide images (available for 1,110 patients), the 40x tiles from the NanoZoomer XR slide images (available for 1,121 patients) and the 10x tiles form the NanoZoomer XR slide images (available for 1,119 patients).



Protocol Fig. 13 | A diagram specifying inclusions and exclusions of patients, slides and slide images from the QUASAR 2 cohort.

The QUASAR 2 cohort is representative for patients eligible for the QUASAR 2 trial. An eligible patient had to satisfy all of the following inclusion criteria (originally described by Kerr et al.[18]):

- Aged 18 years or older.
- Colorectal adenocarcinoma.
- Histologically proven R0 M0 stage III or high-risk stage II colorectal cancer, where high-risk was defined as the presence of one or more of the following adverse prognostic features: stage T4, lymphatic invasion, vascular invasion, peritoneal involvement, poor differentiation and preoperative obstruction or perforation of the primary tumour
- Primary resection between 4 and 10 weeks prior to randomisation.
- World Health Organisation (WHO) Performance Status 0 or 1.
- Life expectancy of at least 5 years when taking into account comorbidities, but excluding cancer risk.

Additionally, an eligible patient could not satisfy any of the following exclusion criteria (originally described by Kerr et al.[18]):

- History of cancer other than treated in-situ carcinoma of the cervix, basal or squamous-cell carcinoma or if the disease-free interval after a previous cancer was greater than 10 years.
- Inflammatory bowel disease and/or active peptic ulcer requiring treatment in the last 2 years.
- Lack of physical integrity of the upper gastrointestinal tract, malabsorption syndrome or inability to take oral medication.
- Moderate or severe renal impairment (creatinine clearance <30 mL/min).
- Any of the following blood abnormalities:
    - Absolute neutrophil count $<1.5x10^9$/L.
    - Platelet count $<100x10^9$/L.
    - Total bilirubin concentration >1.5 times the upper limit of normal (ULN).

    - Alanine aminotransferase, aspartate aminotransferase or alkaline phosphatase concentration >2.5 times the upper limit of normal (ULN).

- Proteinuria >500 mg per 24 hours.
- Previous chemotherapy, immunotherapy or infra-diaphragmatic radiotherapy (including neoadjuvant therapy to the rectum) or patients who are expected to require radiotherapy to these sites within the next 12 months.
- Use of any investigational drug or agent/procedure within 4 weeks of randomisation.
- Chronic use of full-dose anticoagulants, high-dose aspirin (>325 mg/day), anti-platelet drugs or known bleeding diathesis (low-dose aspirin was allowed).
- Concomitant treatment with sorivudine or its chemically related analogues.
- History of uncontrolled seizures, central nervous system disorders or psychiatric precluding informed consent or interfering with compliance for oral drug intake.
- Clinically significant cardiovascular disease, i.e. active or <12 months since e.g. cerebrovascular accident, myocardial infarction, unstable angina, New York Heart Association (NYHA) grade II or greater congestive heart failure, serious cardiac arrhythmia requiring medication or uncontrolled hypertension.
- Known coagulopathy.
- Known allergy to Chinese hamster ovary cell proteins or other recombinant human or humanised antibodies or to any excipients of bevacizumab formulation.
- Women who were pregnant or lactating, or premenopausal women not using contraception.

9 Primary analysis

We predefined a primary analysis of the DoMore v1 marker for each scanner (Aperio AT2 and NanoZoomer XR) in the QUASAR 2 cohort. The selected metric for measuring model performance was the hazard ratio (with 95% confidence interval [CI]) of patients predicted as uncertain prognosis and patients predicted as poor

prognosis relative to patients predicted as good prognosis, where the two hazard ratios (and their corresponding CIs) will be computed by analysing a Cox proportional hazard model with the DoMore v1 marker as the only variable (the DoMore v1 marker will be included as a categorical variable, i.e. the model will consist of the two indicator variables for uncertain prognosis and poor prognosis) and cancer-specific survival as endpoint (Efron's method will be used in case of tied events). The selected test for assessing whether the DoMore v1 marker predicts cancer-specific survival was the two-tailed Mantel-Cox logrank test using significance level 0.05. Time to cancer-specific survival will be computed from date of randomisation to date of cancer-specific death or loss to follow-up. The primary analysis is an unbiased evaluation of the DoMore v1 marker's ability to predict cancer-specific survival in the target population of patients that received adjuvant chemotherapy (specifically capecitabine) and satisfied the eligibility criteria of the QUASAR 2 trial.

10 Secondary analyses

The following secondary analyses were planned in advance of all investigations in the QUASAR 2 cohort that could reveal associations between the slides and clinical outcome, i.e. the recorded cancer-specific survival:

- Repeat the primary analysis for the constituents of the DoMore v1 marker, i.e. the two dichotomous 10x and 40x ensemble markers of the DoMore v1 network. Note that since these are dichotomous markers, there will only be one hazard ratio for each of them, i.e. that of patients predicted as poor prognosis relative to patients predicted as good prognosis.

- Repeat the primary analysis (of the DoMore v1 marker) for stage II and III patients separately.

- Include the DoMore v1 marker evaluated on Aperio AT2 slide images as a categorical variable (as in the primary analysis) in a multivariable model together with relevant markers that is available at the time of analysis and is significant in univariable analysis of cancer-specific survival (defined as in the primary analysis) in the QUASAR 2 cohort. Analyse the model with the same endpoint definition as in the primary analysis. Compute the hazard ratio (with 95% CI) and corresponding *P* value of patients predicted as uncertain prognosis and patients predicted as poor prognosis relative to patients predicted as good prognosis using Cox proportional hazard model and Wald X^2 test when analysing only patients with complete data for all variables included in the multivariable model. Current candidate markers are:

    - Pathological N stage (with categories N0, N1 and N2). Note that this marker incorporates the pathological stage, i.e. stage II or stage III.
    - Pathological T stage (with categories 1, 2, 3 and 4).
    - Age at randomisation (continuous on linear scale).
    - Sex (with categories *Female* and *Male*).
    - Histological grade (with categories 1, 2 and 3).
    - Location. (with categories *Proximal colon*, *Distal colon* and *Rectum* where *Proximal colon* includes the cecum through the transverse colon and *Distal colon* includes the left flexure through the rectosigmoid flexure).
    - Venous vascular invasion (with categories *No* and *Yes*).
    - Lymphatic invasion (with categories *No* and *Yes*).
    - MSI (with categories *Unstable* and *Stable*).
    - BRAF (if available).
    - KRAS (if available).

- Repeat the primary analysis for the Inception v3 marker.
- Repeat the primary analysis for the constituents of the Inception v3 marker, i.e. the two dichotomous 10x and 40x ensemble markers of the Inception v3 network. Again note that since these are dichotomous markers, there will only be one hazard ratio for each of them, i.e. that of patients predicted as poor prognosis relative to patients predicted as good prognosis.

References

1. Bondi, J., *et al.* Expression and gene amplification of primary (A, B1, D1, D3, and E) and secondary (C and H) cyclins in colon adenocarcinomas and correlation with patient outcome. *J Clin Pathol* **58**, 509-514 (2005).

2. Danielsen, H.E., *et al.* Prognostic markers for colorectal cancer: estimating ploidy and stroma. *Ann Oncol* **29**, 616-623 (2018).
3. Petersen, V.C., Baxter, K.J., Love, S.B. & Shepherd, N.A. Identification of objective pathological prognostic determinants and models of prognosis in Dukes' B colon cancer. *Gut* **51**, 65-69 (2002).
4. Mitchard, J.R., Love, S.B., Baxter, K.J. & Shepherd, N.A. How important is peritoneal involvement in rectal cancer? A prospective study of 331 cases. *Histopathology* **57**, 671-679 (2010).
5. Kerr, D.J., *et al.* Rofecoxib and Cardiovascular Adverse Events in Adjuvant Treatment of Colorectal Cancer. *N Engl J Med* **357**, 360-369 (2007).
6. Midgley, R.S., *et al.* Phase III Randomized Trial Assessing Rofecoxib in the Adjuvant Setting of Colorectal Cancer: Final Results of the VICTOR Trial. *J Clin Oncol* **28**, 4575-4580 (2010).
7. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans Pattern Anal Mach Intell* **40**, 834-848 (2018).
8. Krähenbühl, P. & Koltun, V. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. in *Adv Neural Inf Process Syst*, Vol. 24 109-117 (2011).
9. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. in *Proc 13th Int Conf Artif Intell Stat*, Vol. 9 249-256 (2010).
10. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. in *IEEE Conf Comput Vis Pattern Recognit* 4510-4520 (2018).
11. Glorot, X., Bordes, A. & Bengio, Y. Deep Sparse Rectifier Neural Networks. in *Proc 14th Int Conf Artif Intell Stat*, Vol. 15 315-323 (2011).
12. Krizhevsky, A. Convolutional Deep Belief Networks on CIFAR-10. Available from: https://www.cs.toronto.edu/~kriz/conv-cifar10-aug2010.pdf. (2010).
13. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. in *Proc 32nd Int Conf Mach Learn*, Vol. 37 448-456 (2015).
14. Kraus, O.Z., Ba, J.L. & Frey, B.J. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics* **32**, i52-i59 (2016).
15. Kingma, D.P. & Ba, J. Adam: A Method for Stochastic Optimization. Available from: https://arxiv.org/abs/1412.6980. (2015).

16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception Architecture for Computer Vision. in *Proc 2016 IEEE Conf Comput Vis Pattern Recognit* 2818-2826 (2016).

17. Harrell, F.E., Jr, Califf, R.M., Pryor, D.B., Lee, K.L. & Rosati, R.A. Evaluating the yield of medical tests. *J Am Med Assoc* **247**, 2543-2546 (1982).

18. Kerr, R.S., *et al.* Adjuvant capecitabine plus bevacizumab versus capecitabine alone in patients with colorectal cancer (QUASAR 2): an open-label, randomised phase 3 trial. *Lancet Oncol* **17**, 1543-1557 (2016).